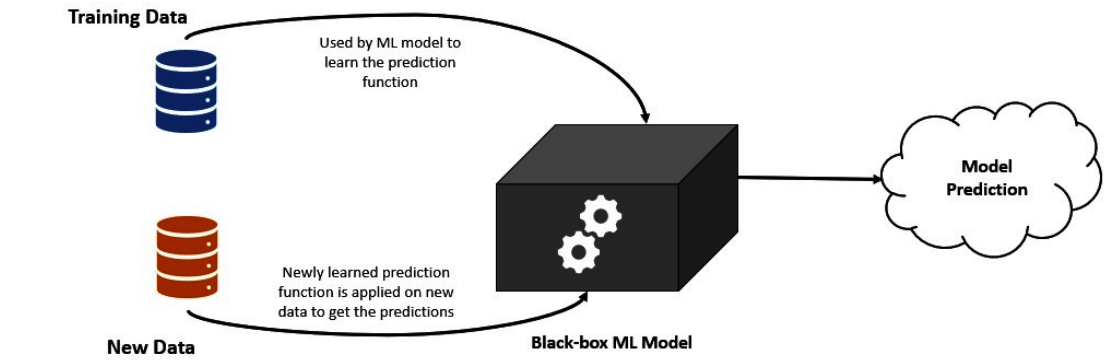
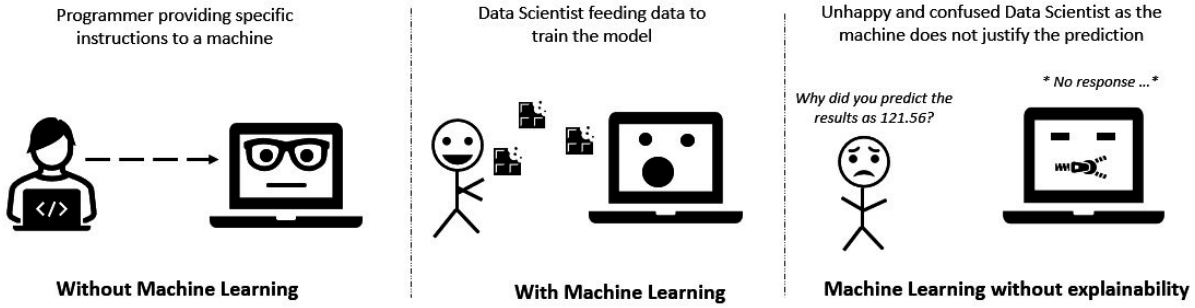
Chapter 1: Foundational Concepts of Explainability Techniques





Overcoming Racial Bias In AI Systems And Starlingly Even In AI Self-Driving Cars

Racial bias in a medical algorithm favors white patients over sicker black patients

AI expert calls for end to UK use of 'racially biased' algorithms

AI Bias Could Put Women's Lives At Risk - A Challenge For Regulators

Gender bias in AI: building fairer algorithms

Bias in AI: A problem recognized but still unresolved

Amazon, Apple, Google, IBM, and Microsoft worse at transcribing black people's voices than white people's with AI voice recognition, study finds

Millions of black people affected by racial bias in health-care algorithms

Study reveals rampant racism in decisions-making software used by US hospitals and highlights ways to correct it.

When it Comes to Gorillas, Google Photos Blind

Google promised a fix after its photo-categorization software labeled black people as gorillas in 2015. More than two years later, it hasn't found one.

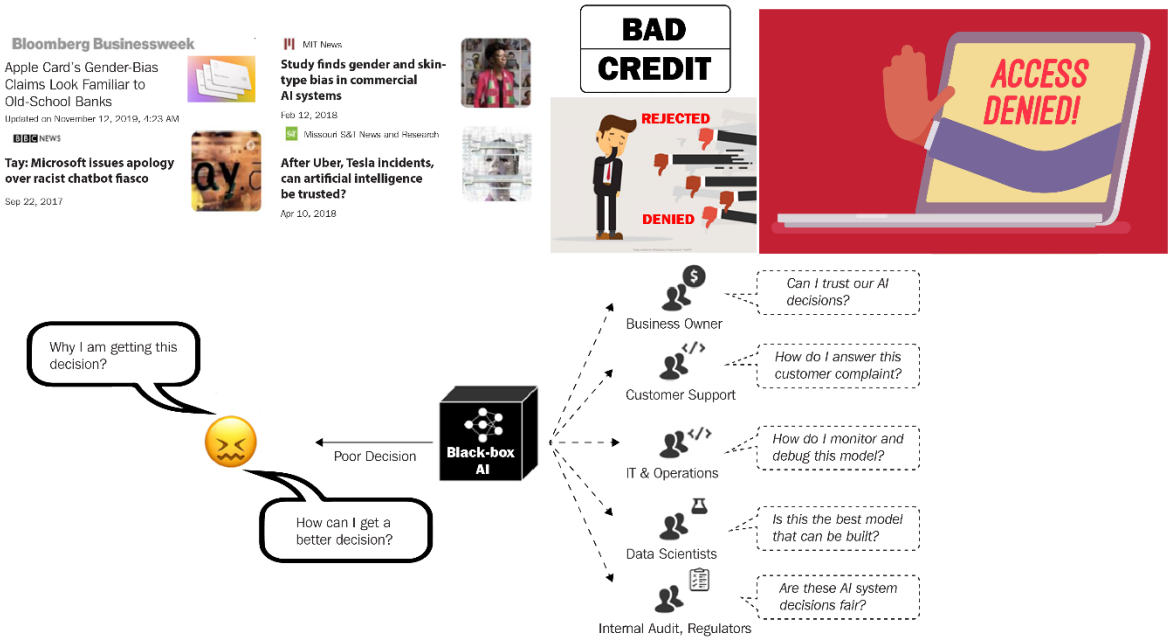
The Week in Tech: Algorithmic Bias Is Bad. Uncovering It Is Good.

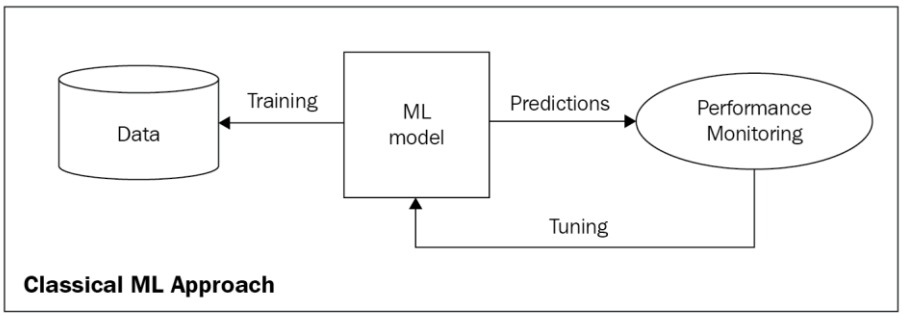
Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech

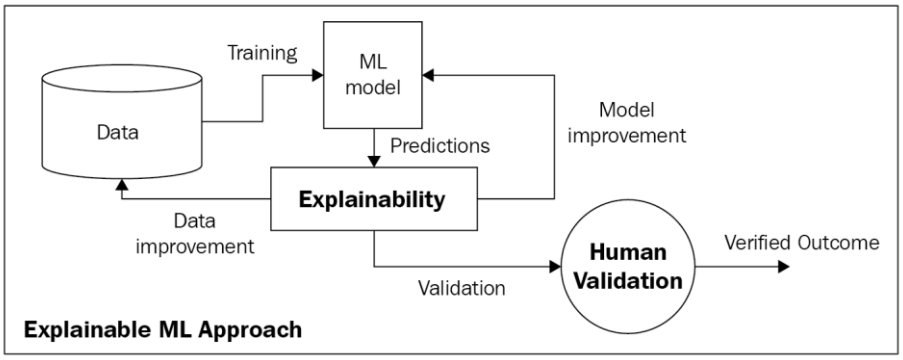
Artificial Intelligence has a gender bias problem - just ask Siri

The Best Algorithms Struggle to Recognize Black Faces Equally

US government tests find even top-performing facial recognition systems misidentify blacks at rates fives to 10 times higher than they do whites.


Bloomberg Businessweek
Apple Card's Gender-Bias Claims Look Familiar to Old-School Banks
Updated on November 12, 2019, 4:23 AM
BBC NEWS
Tay: Microsoft issues apology over racist chatbot fiasco
Sep 22, 2017
MIT News
Study finds gender and skin-type bias in commercial AI systems
Feb 12, 2018
Missouri S&T News and Research
After Uber, Tesla incidents, can artificial intelligence be trusted?
Apr 10, 2018
BAD CREDIT
REJECTED
DENIED
ACCESS DENIED!
Why I am getting this decision?
How can I get a better decision?
Poor Decision
Black-box AI
Business Owner
Can I trust our AI decisions?
Customer Support
How do I answer this customer complaint?
IT & Operations
How do I monitor and debug this model?
Data Scientists
Is this the best model that can be built?
Internal Audit, Regulators
Are these AI system decisions fair?


Data
Training
ML model
Predictions
Performance Monitoring
Tuning
Classical ML Approach


Data
Training
ML model
Predictions
Model improvement
Explainability
Data improvement
Validation
Human Validation
Verified Outcome
Explainable ML Approach

"

Companies should commit to ensuring systems that could fall under GDPR, including AI, will be complaint. The threat of **sizeable fines of €20 million or 4% of global turnover** provides a sharp incentive.

Article 22 of GDPR empowers individuals with the **right to demand an explanation of how an AI system made a decision that affects them**.

"

- European Commision



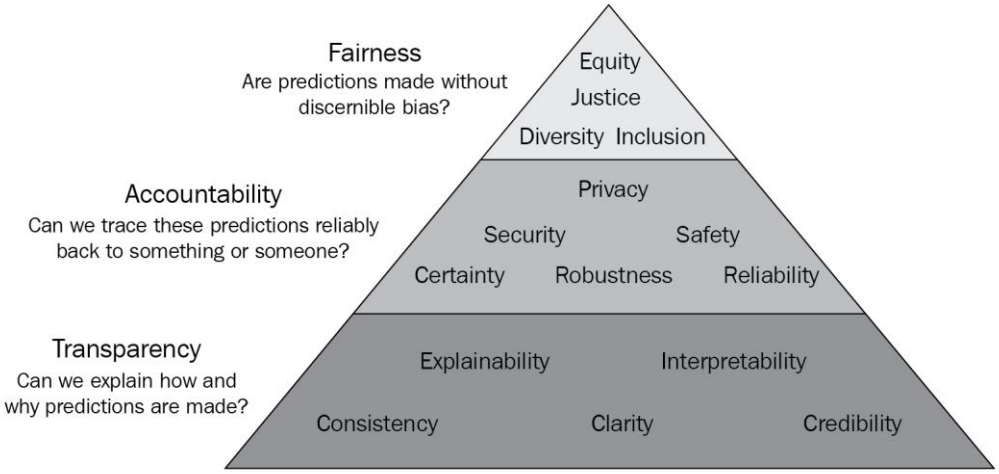
Andrus Ansip
@Ansip_EU

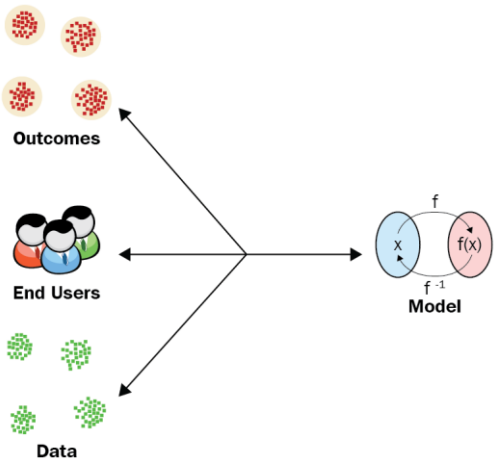
You have the right to be informed about an automated decision and ask for a human being to review it, for example if your online credit application is refused. #EUdataP #GDPR #AI #digitalrights #EUandMe europa.eu/!nN77Dd

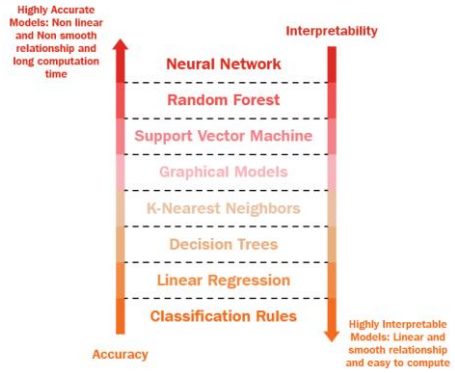


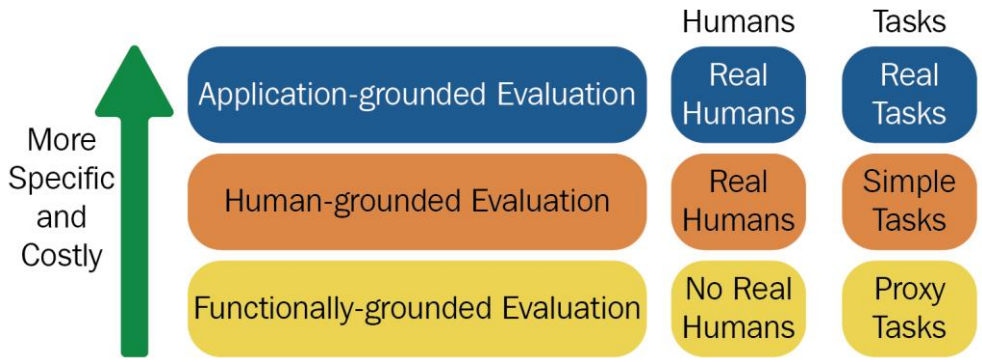
VP, European Commision

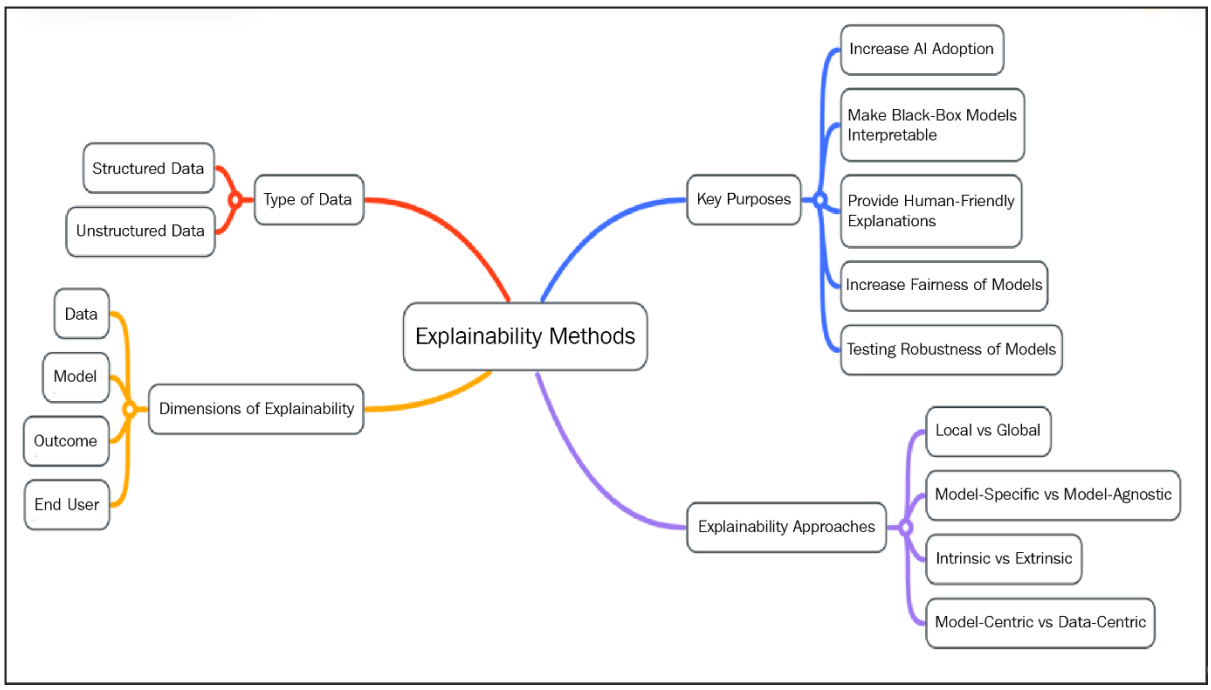
Regulatory Acts	Descriptions
GDPR	Article 22 empowers individuals with the right to demand an explaination of the decision-making process of automated systems
California Consumer Privacy Act 2019	Requires companies to align the process of collection, storage, and sharing of personal data with the new requirements of January 1, 2020
Algorithmic Accountability Act 2019	Mandates organizations to provide assessments of the risks of having automated decision systems to privacy, security, inaccurate, unfair, biased, or discriminatory outcomes impacting consumers.
Washington Bill 1655	Introduces measures for the use of automated decision systems to protect consumers, improve transparency and create more market predictability.
Illinois House Bill 3415	Establishes guidelines for not including information related to applicant race or zip code for predictive data analytics for the purpose of hiring or financial services
Massachusetts Bill H.2701	Establishes guidelines on automated decision-making, transparency, fairness, and individual rights.



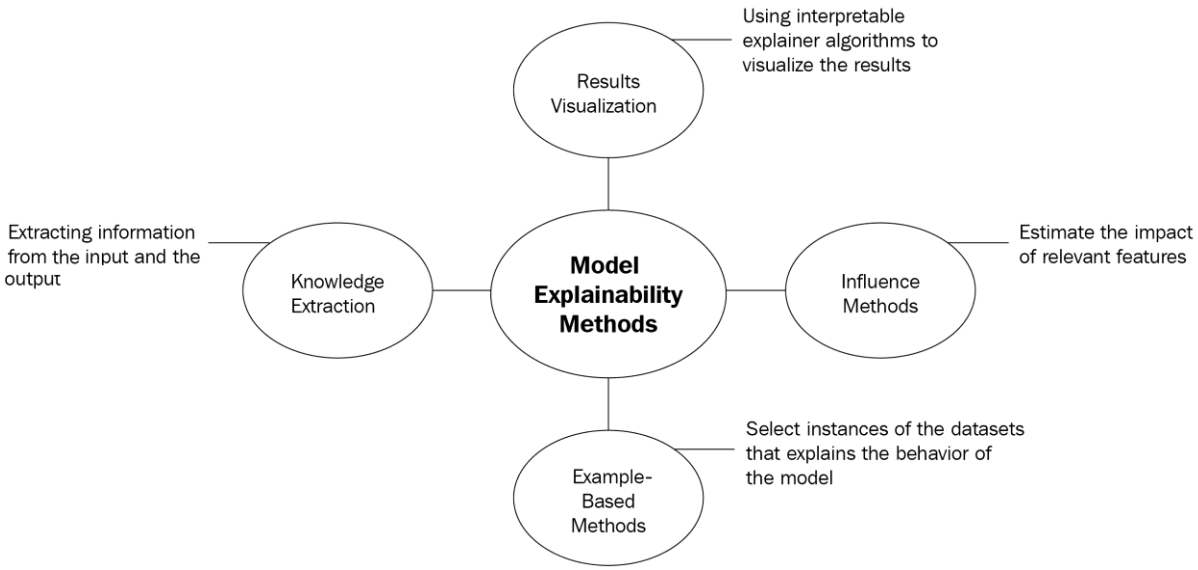

Outcomes
End Users
Data
f
x
f(x)
f -1
Model


Highly Accurate Models: Non linear and Non smooth relationship and long computation time
Interpretability
Neural Network
Random Forest
Support Vector Machine
Graphical Models
K-Nearest Neighbors
Decision Trees
Linear Regression
Classification Rules
Accuracy
Highly Interpretable Models: Linear and smooth relationship and easy to compute


Humans
Tasks
More Specific and Costly
Application-grounded Evaluation
Real Humans
Real Tasks
Human-grounded Evaluation
Real Humans
Simple Tasks
Functionally-grounded Evaluation
No Real Humans
Proxy Tasks


Explainability Methods
Type of Data
Structured Data
Unstructured Data
Dimensions of Explainability
Data
Model
Outcome
End User
Key Purposes
Increase AI Adoption
Make Black-Box Models Interpretable
Provide Human-Friendly Explanations
Increase Fairness of Models
Testing Robustness of Models
Explainability Approaches
Local vs Global
Model-Specific vs Model-Agnostic
Intrinsic vs Extrinsic
Model-Centric vs Data-Centric

Chapter 2: Model Explainability Methods

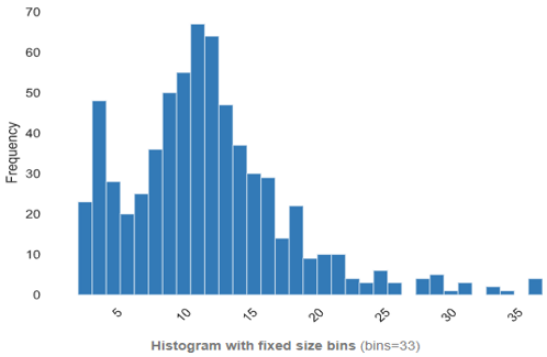


Dataset statistics

Number of variables	15
Number of observations	660
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%

Variable types

Categorical	5
Numeric	10



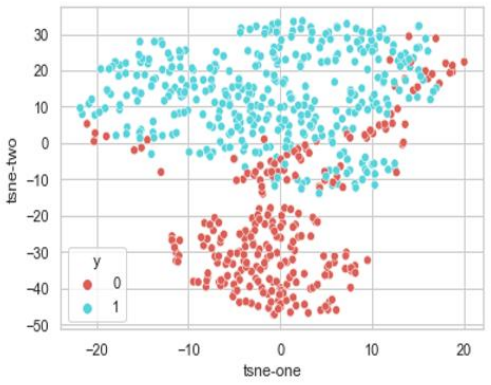


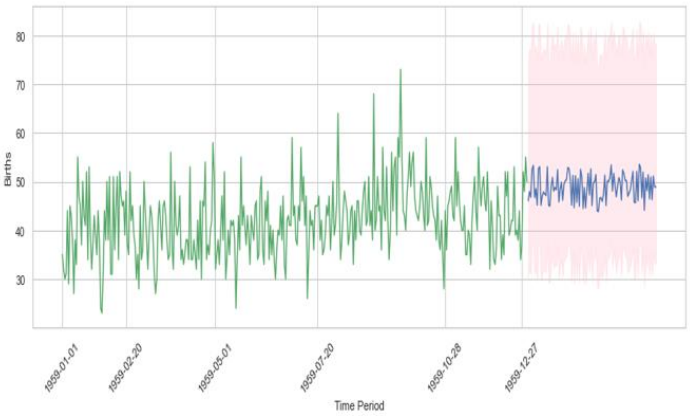
Quantile statistics

Minimum	2
5-th percentile	4
Q1	8
median	11
Q3	14
95-th percentile	23
Maximum	37
Range	35
Interquartile range (IQR)	6

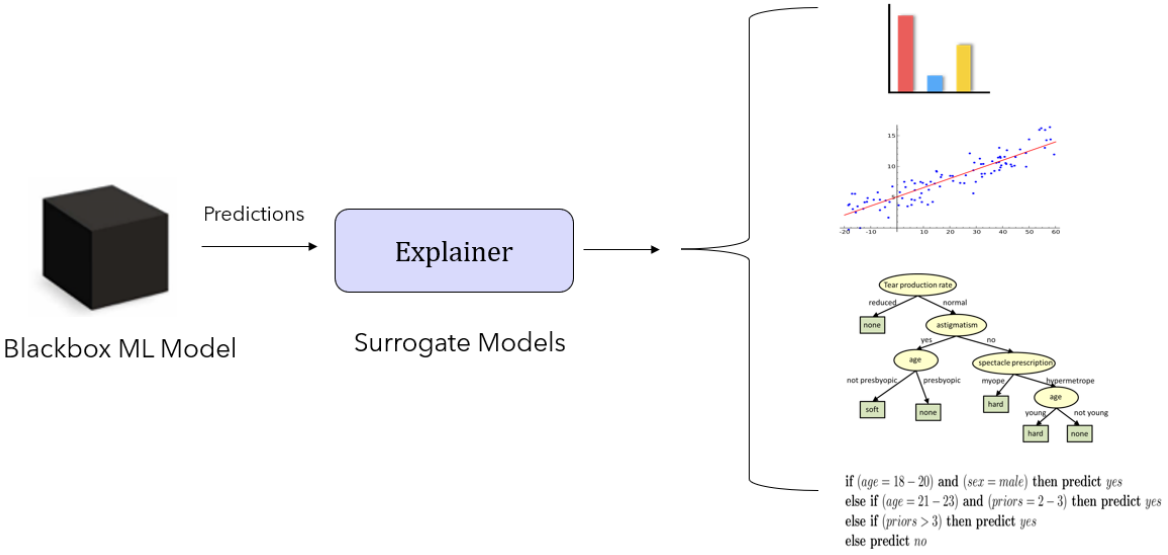
Descriptive statistics

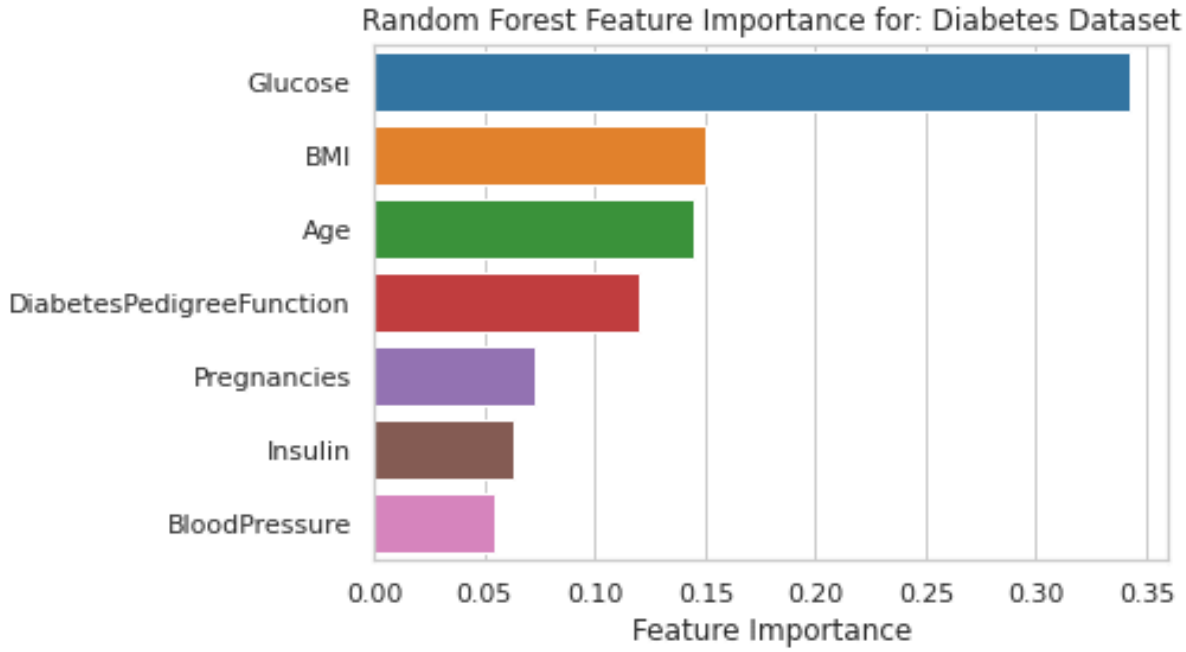
Standard deviation	5.98245392
Coefficient of variation (CV)	0.507639443
Kurtosis	2.330457684
Mean	11.78484848
Median Absolute Deviation (MAD)	3
Skewness	1.181334957
Sum	7778
Variance	35.78975491
Monotonicity	Not monotonic





Algorithm	Relationship	Type of ML problem
Decision trees	Interactions and monotonicity (to some extent)	Classification and regression
RuleFit	Linearity and interactions	Classification and regression
Linear regression	Linearity and monotonicity	Regression
Logistic regression	Monotonicity	Classification
Naïve Bayes	Monotonicity	Classification





Sensitivity Analysis

A quantitative process which accounts for uncertainty in forecasts by changing assumptions made about variables important.

- Questioning the reliability of variable assumptions.
- Predicts what will happen if assumptions are changed.
- The significance of changing variable assumptions.

Factors Influencing Forecasts

- Rivals Change Actions
- Exchange Rate Changes
- Interest Rate Changes
- Unexpected Weather Events

These events will impact a business's strategic variables and influence the validity of forecasts.

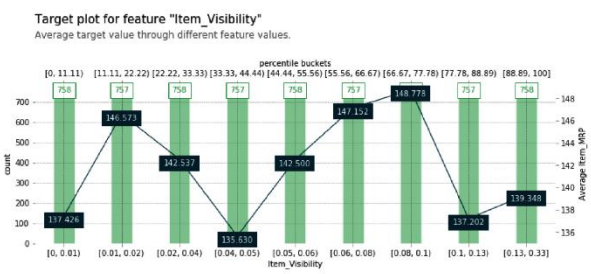
Performing Sensitivity Analysis

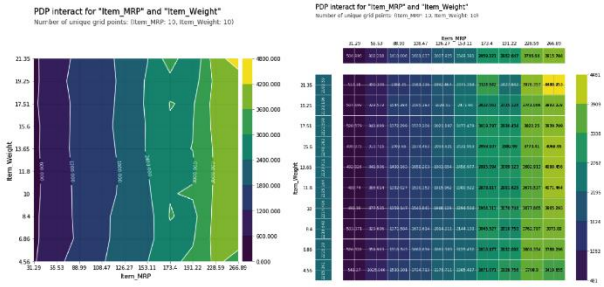
	Negative Outcome	Expected Outcome	Positive Outcome
Price	£15	£20	£25
Variable Costs	£9	£7	£5
Fixed Costs	£5,000	£4,000	£3,000
Forecasted Sales	1,000	2,000	3,000
	Business takes a negative perspective of the business	The business expectation of what will happen.	Business takes a positive perspective of the business
	Lower Profits	Expected Profits	Higher Profits
	£1,000	£22,000	£57,000

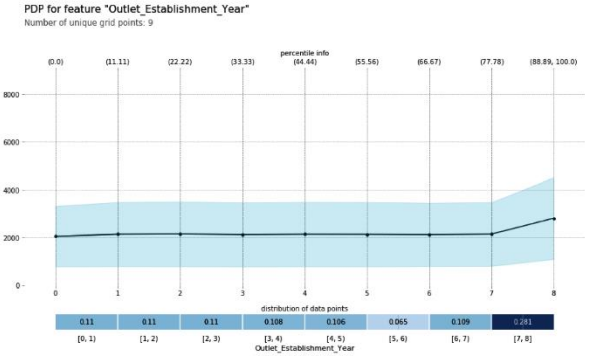
Spectrum of business scenarios forecast

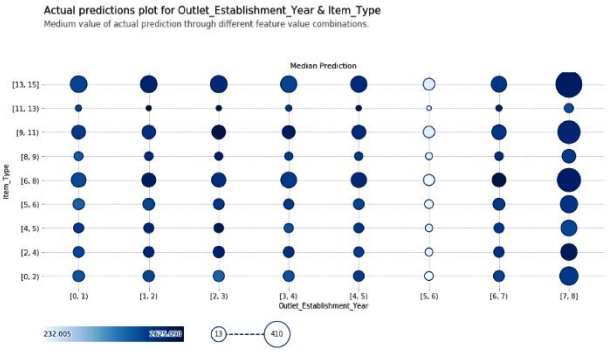
Evaluating Sensitivity Analysis

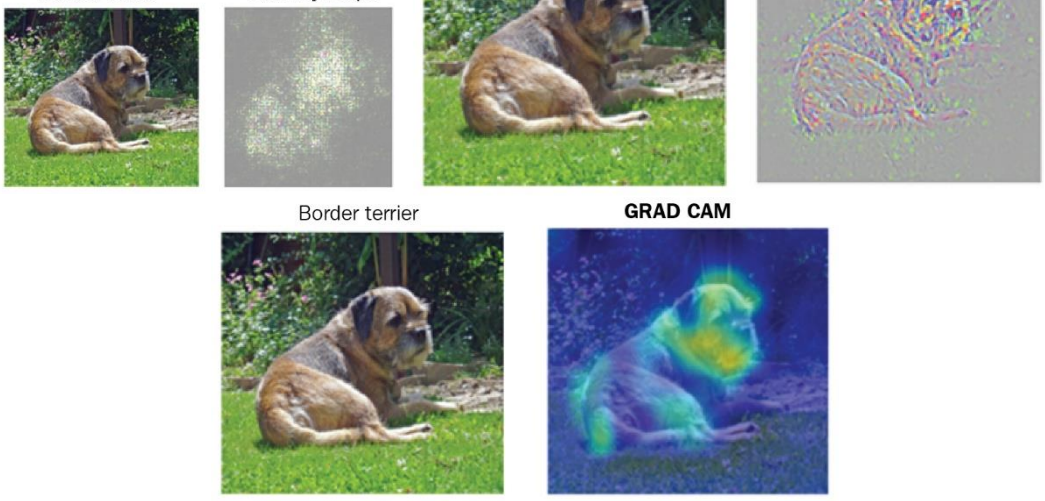
- Allows managers to focus on the variables that they can influence.
- In preparation for the worst-case scenario.
- Important to reassure investors that the business is sufficiently prepared.
- Sensitivity analysis is a system driven by data...
  - Is that data reliable?
  - Is that data accurate?
  - Is that data relevant?
  - Are there any other interventions?

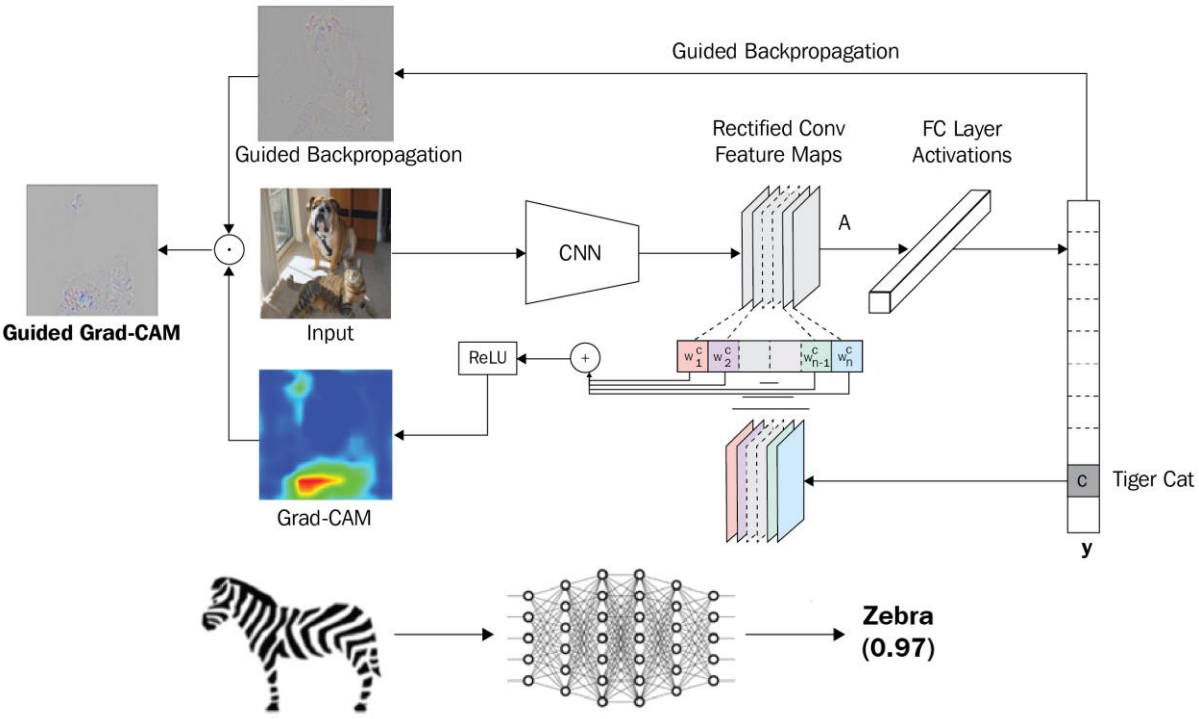

Target plot for feature "Item_Visibility"
Average target value through different feature values.
percentile buckets
[0, 11.11) [11.11, 22.22) [22.22, 33.33) [33.33, 44.44) [44.44, 55.56) [55.56, 66.67) [66.67, 77.78) [77.78, 88.89) [88.89, 100]
758 757 758 757 758 757 758 757 758
137.426 146.573 142.537 135.630 142.500 147.152 148.778 137.202 139.348
count
Average Item_MRP
[0, 0.01) [0.01, 0.02) [0.02, 0.04) [0.04, 0.05) [0.05, 0.06) [0.06, 0.08) [0.08, 0.1) [0.1, 0.13) [0.13, 0.33]
Item_Visibility

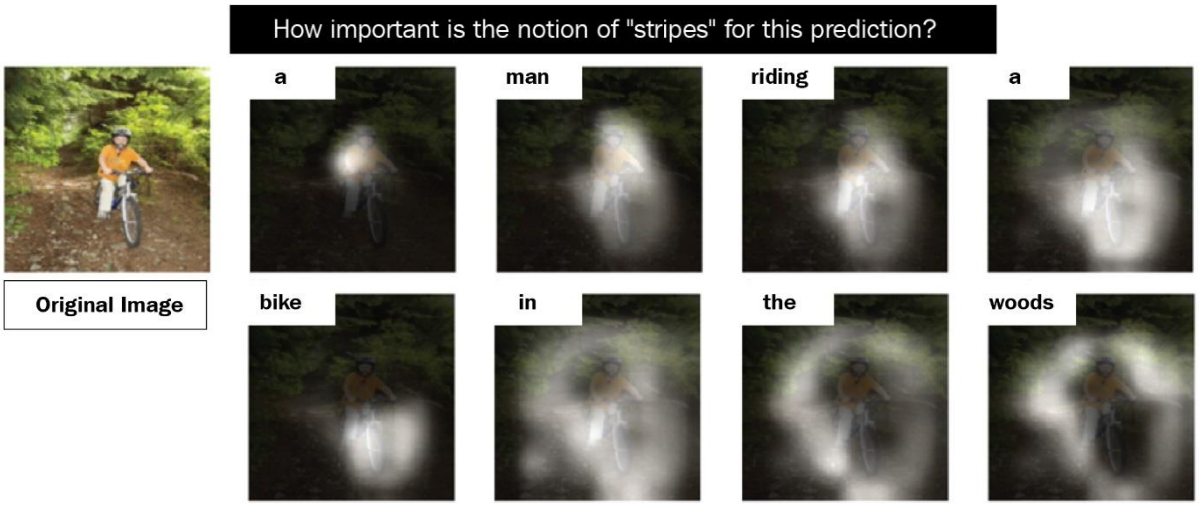

PDP interact for "Item_MRP" and "Item_Weight"
Number of unique grid points: (Item_MRP: 10, Item_Weight: 10)
Item_Weight
Item_MRP
PDP interact for "Item_MRP" and "Item_Weight"

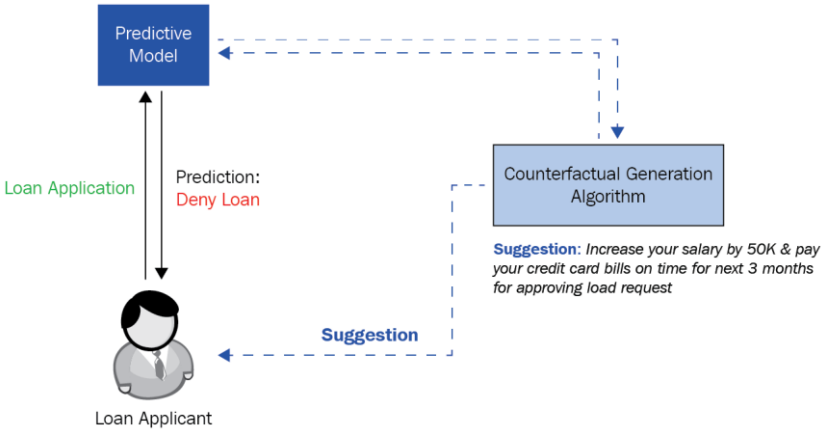

PDP for feature "Outlet_Establishment_Year"
Number of unique grid points: 9
percentile info
(0.0) (11.11) (22.22) (33.33) (44.44) (55.56) (66.67) (77.78) (88.89, 100.0)
distribution of data points
0.11 0.11 0.11 0.108 0.106 0.065 0.109 0.281
[0, 1) [1, 2) [2, 3) [3, 4) [4, 5) [5, 6) [6, 7) [7, 8)
Outlet_Establishment_Year

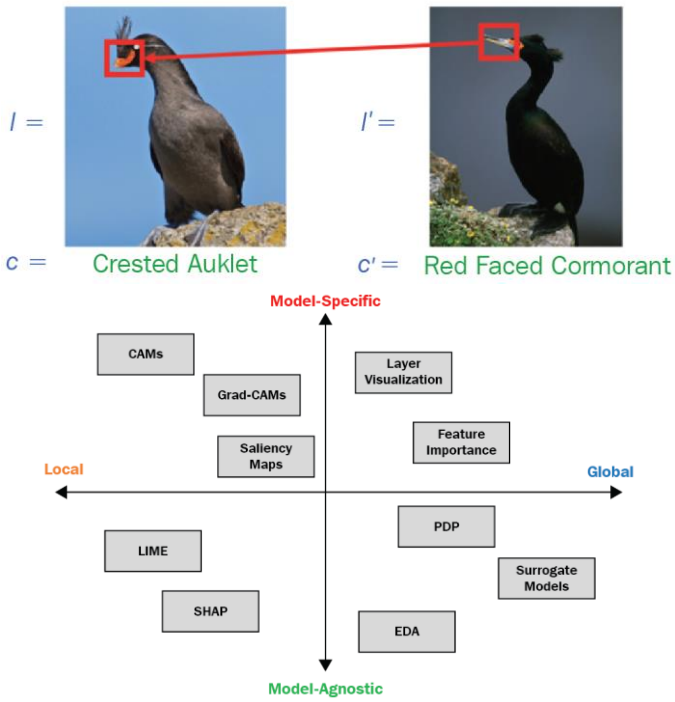

Actual predictions plot for Outlet_Establishment_Year & Item_Type
Medium value of actual prediction through different feature value combinations.
Median Prediction
Item_Type
[13, 15]
[11, 13)
[9, 11)
[8, 9)
[6, 8)
[5, 6)
[4, 5)
[2, 4)
[0, 2)
[0, 1) [1, 2) [2, 3) [3, 4) [4, 5) [5, 6) [6, 7) [7, 8)
Outlet_Establishment_Year
232.005
13
410


Border terrier
Saliency Maps
Border terrier
Guided Backprop
Border terrier
GRAD CAM


Guided Backpropagation
Guided Backpropagation
Rectified Conv Feature Maps
FC Layer Activations
Guided Grad-CAM
Input
CNN
A
ReLU
Grad-CAM
C
Tiger Cat
y
Zebra
(0.97)


How important is the notion of "stripes" for this prediction?
Original Image
a
man
riding
a
bike
in
the
woods

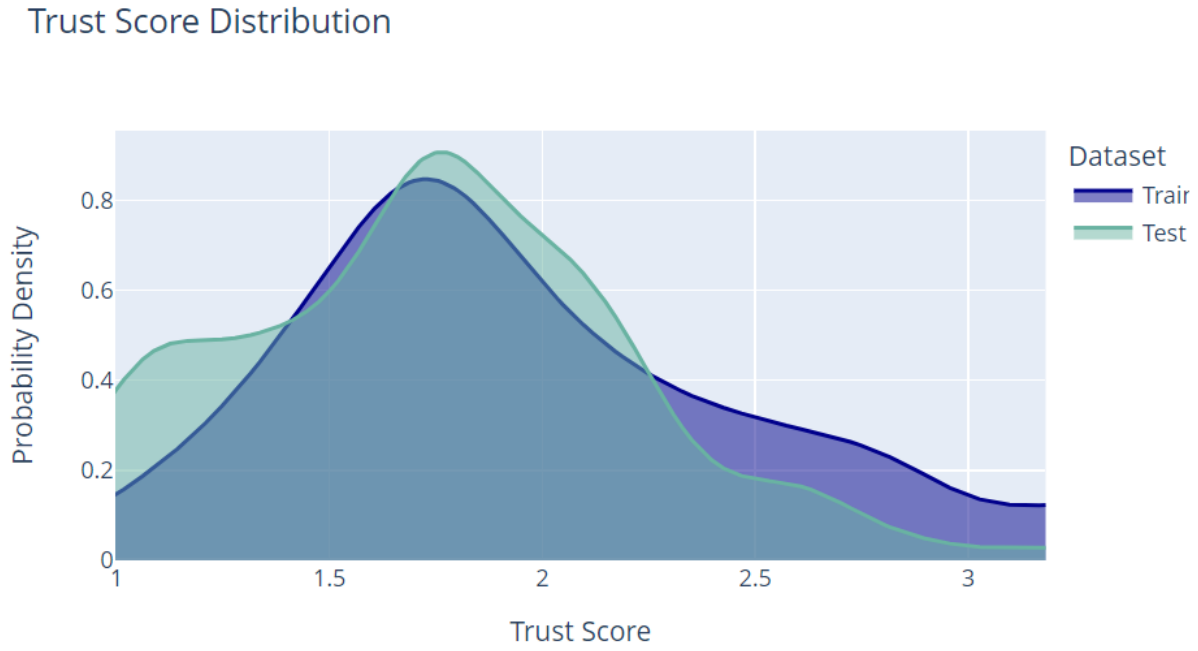

Predictive Model
Loan Application
Prediction:
Deny Loan
Counterfactual Generation Algorithm
Suggestion: Increase your salary by 50K & pay your credit card bills on time for next 3 months for approving load request
Suggestion
Loan Applicant

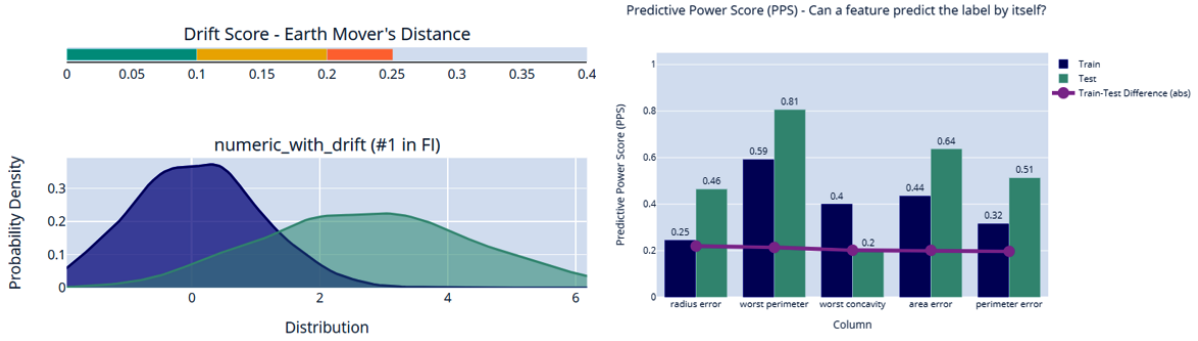

I =
I' =
c =
Crested Auklet
c' =
Red Faced Cormorant
Model-Specific
CAMs
Grad-CAMs
Saliency Maps
Layer Visualization
Feature Importance
Local
Global
PDP
LIME
Surrogate Models
SHAP
EDA
Model-Agnostic

Chapter 3: Data-Centric Approaches

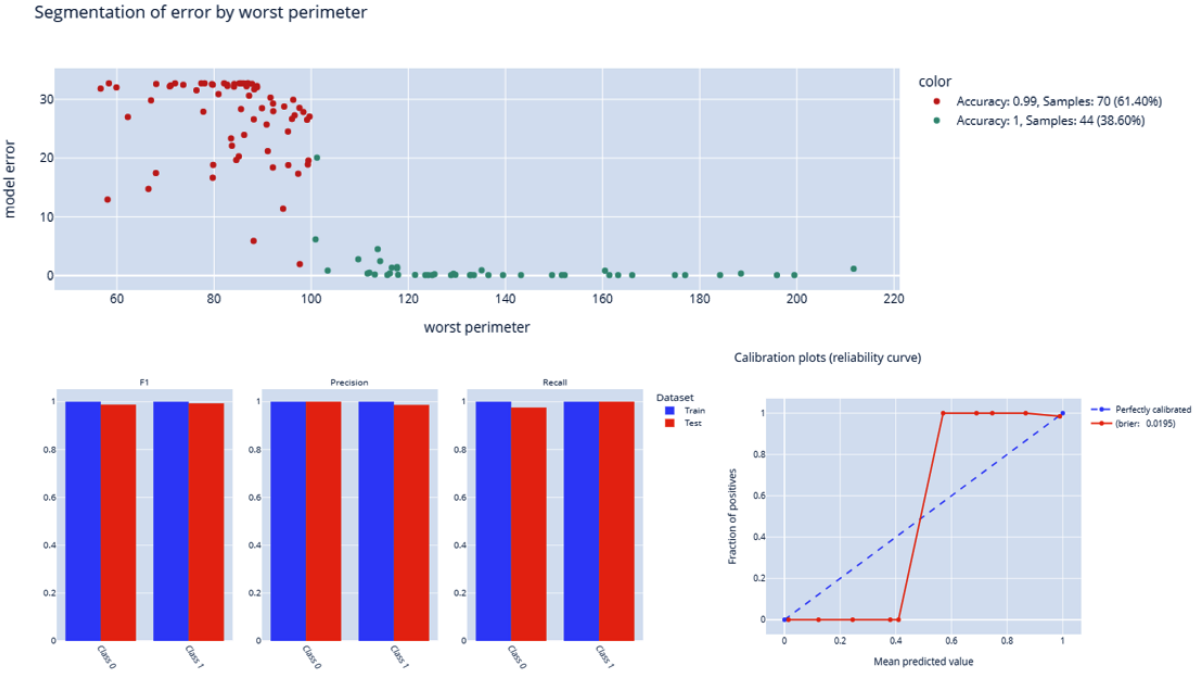
Status	Check	Condition
✓	Single Value in Column - Test Dataset	Does not contain only a single value for all columns
✓	Mixed Nulls - Test Dataset	Not more than 1 different null types for all columns
✓	Mixed Data Types - Test Dataset	Rare data types in all columns are either more than 10.00% or less than 1.00% of the data
✓	String Mismatch - Test Dataset	No string variants for all columns
✓	Data Duplicates - Test Dataset	Duplicate data is not greater than 0%
✓	String Length Out Of Bounds - Test Dataset	Ratio of outliers not greater than 0% string length outliers for all columns
✓	Special Characters - Test Dataset	Ratio of entirely special character samples not greater than 0.10% for all columns
✓	Label Ambiguity - Test Dataset	Ambiguous sample ratio is not greater than 0%

Class	Mean_feat1	Median_feat1	AvgVar_feat1	Mean_feat2	Median_feat2	AvgVar_feat2
0	34.5	37.0	-3.5	128.0	103.5	4.0
1	23.8	23.8	7.8	73.9	102.8	2.2
2	49.0	40	-2.3	101.5	101.5	-1.8





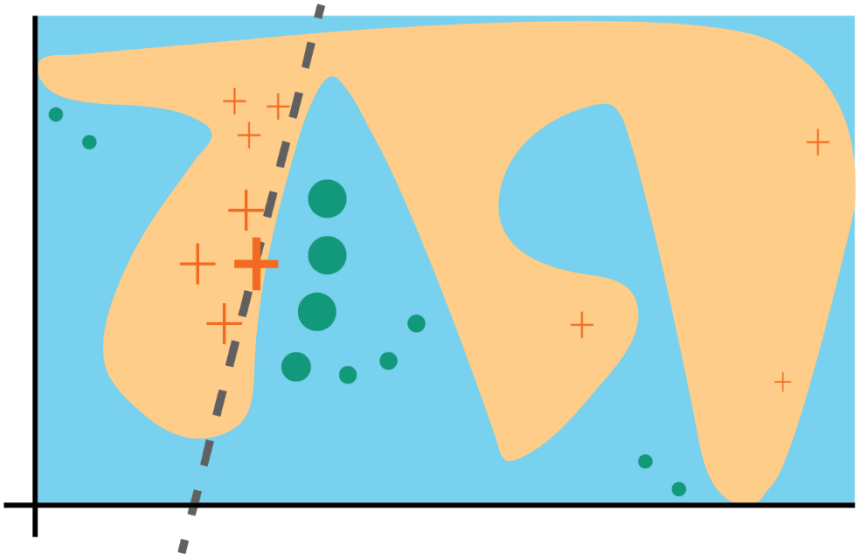

Original. Predicted class: Indian_elephant
FGSM Sample. Predicted class: tusker
C&W Sample. Predicted class: tusker
PGD Sample. Predicted class: triceratops
Original. Predicted: Indian_elephant
Smoothened FGSM Sample. Predicted : tusker
Smoothened C&W Sample. Predicted : tusker
Smoothened PGD Sample. Predicted : Indian_elephant

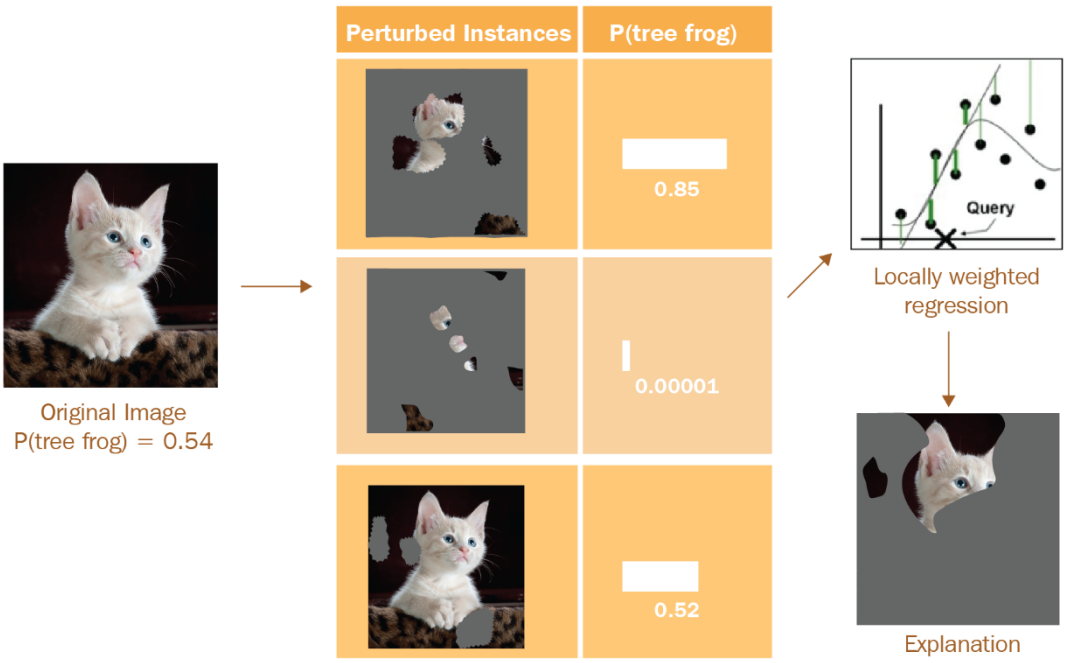

Segmentation of error by worst perimeter
model error
worst perimeter
color
Accuracy: 0.99, Samples: 70 (61.40%)
Accuracy: 1, Samples: 44 (38.60%)
F1
Precision
Recall
Dataset
Train
Test
Class 0
Class 1
Calibration plots (reliability curve)
Fraction of positives
Mean predicted value
Perfectly calibrated
(brier: 0.0195)

Chapter 4: LIME for Model Interpretability





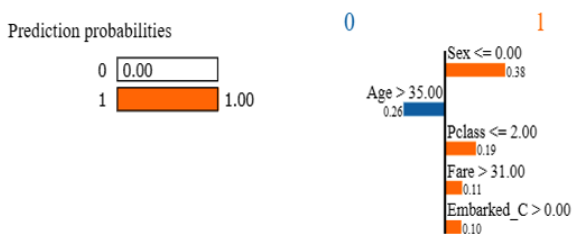


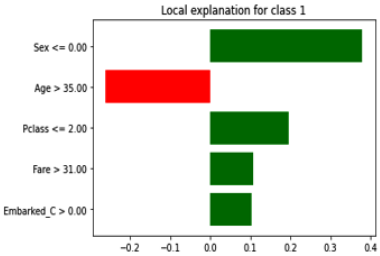


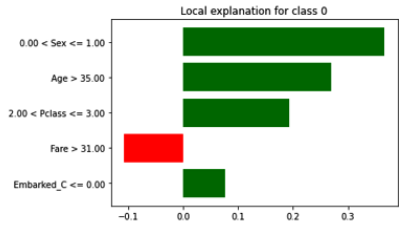
PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q
7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	NaN	S
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	NaN	S
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0708	NaN	C
11	1	3	Sandstrom, Miss. Marguerite Rut	female	4.0	1	1	PP 9549	16.7000	G6	S
12	1	1	Bonnell, Miss. Elizabeth	female	58.0	0	0	113783	26.5500	C103	S
13	0	3	Saundercock, Mr. William Henry	male	20.0	0	0	A/5. 2151	8.0500	NaN	S
14	0	3	Andersson, Mr. Anders Johan	male	39.0	1	5	347082	31.2750	NaN	S
15	0	3	Vestrom, Miss. Hulda Amanda Adolfina	female	14.0	0	0	350406	7.8542	NaN	S

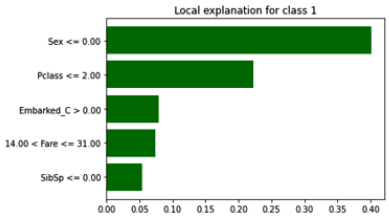
Variable	Description	Details
survival	Survival	0 = No; 1 = Yes
pclass	Passenger Class	1 = 1st; 2 = 2nd; 3 = 3rd
name	First and Last Name	
sex	Sex	
age	Age	
sibsp	Number of Siblings/Spouses Aboard	
parch	Number of Parents/Children Aboard	
ticket	Ticket Number	
fare	Passenger Fare	
cabin	Cabin	
embarked	Port of Embarkation	C = Cherbourg; Q = Queenstown; S = Southampton

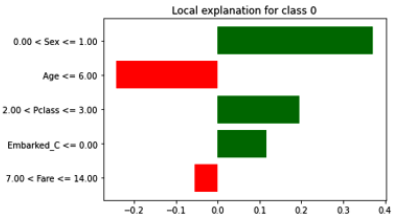
Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked_0	Embarked_C	Embarked_Q	Embarked_S
0	3	1	22.0	1	0	7.2500	0	0	0	1
1	1	0	38.0	1	0	71.2833	0	1	0	0
1	3	0	26.0	0	0	7.9250	0	0	0	1
1	1	0	35.0	1	0	53.1000	0	0	0	1
0	3	1	35.0	0	0	8.0500	0	0	0	1

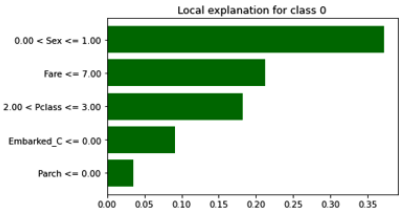


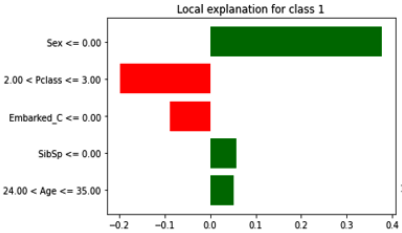


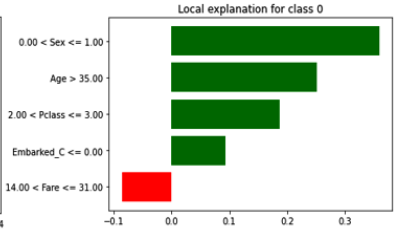




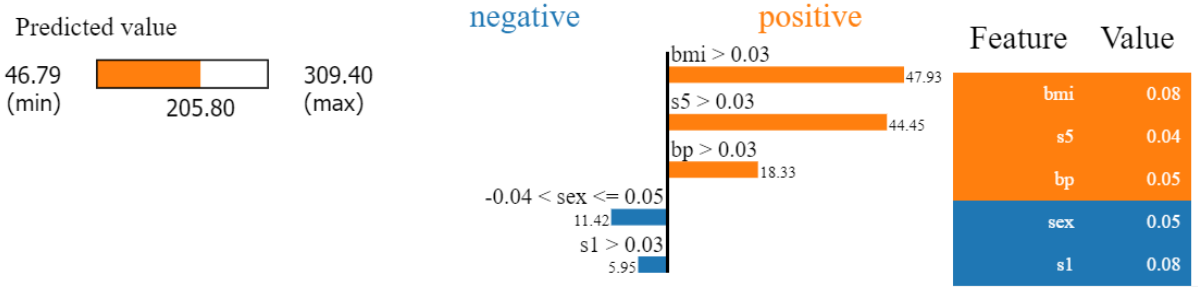


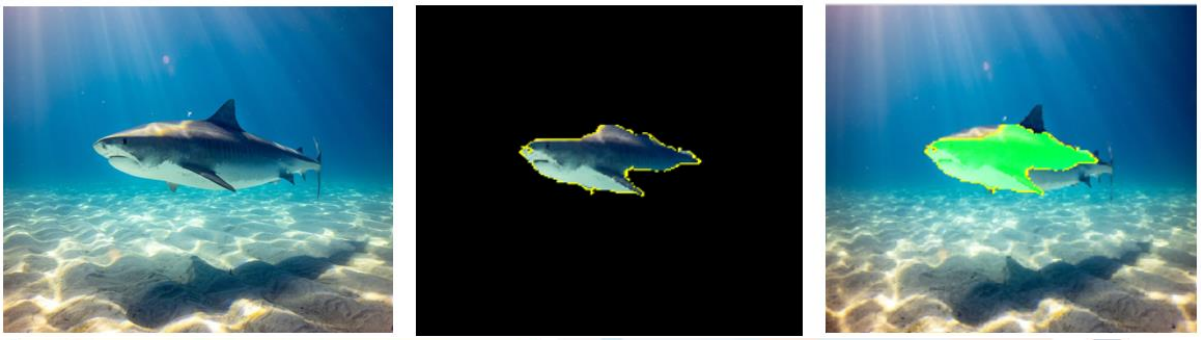


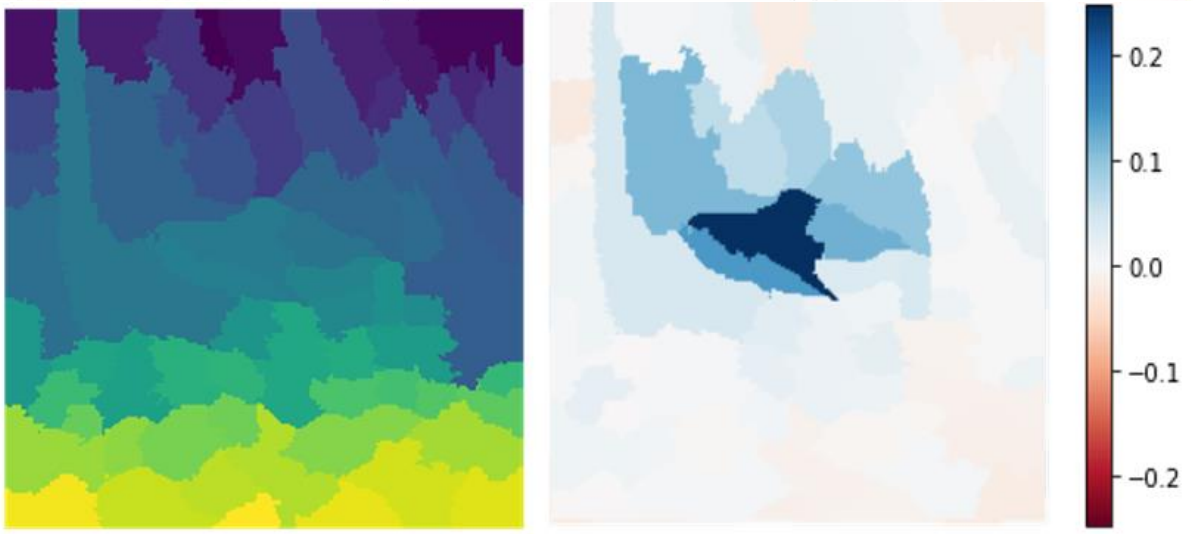


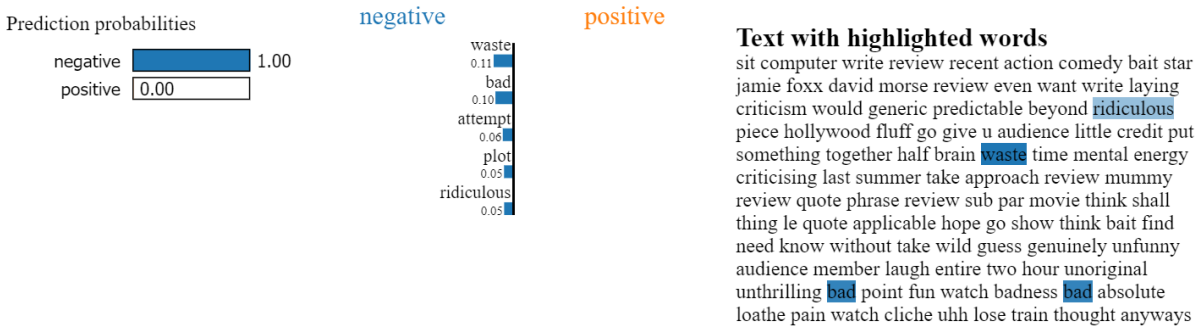


Chapter 5: Practical Exposure to Using LIME in ML





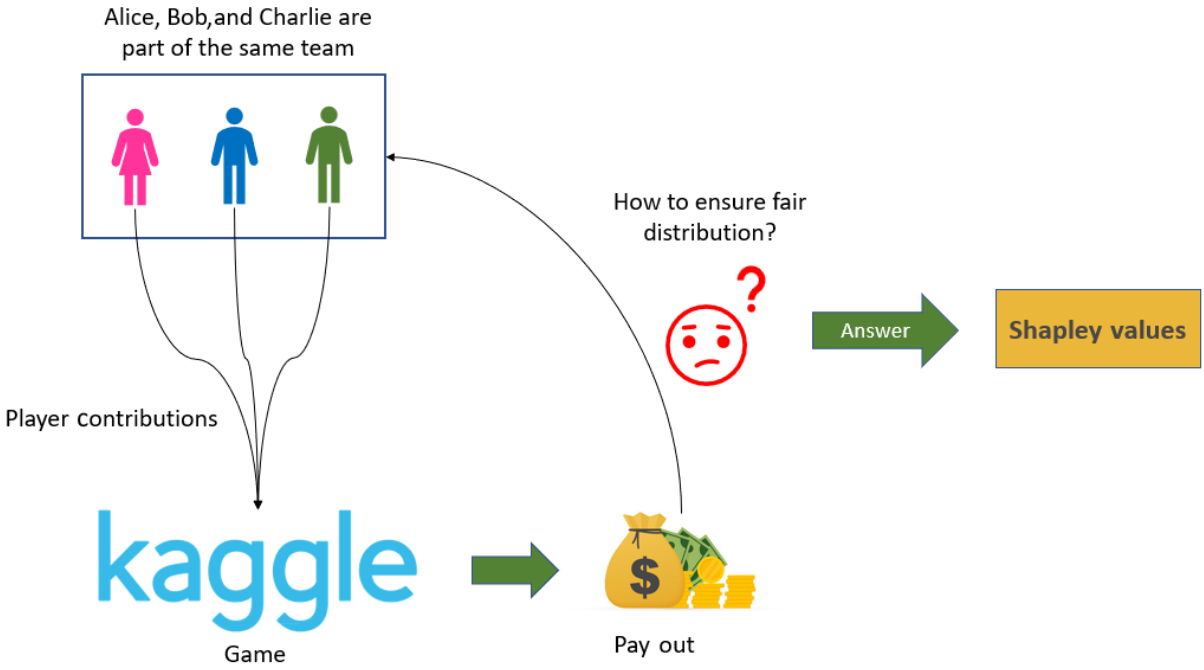




Text with highlighted words

sit computer write review recent action comedy bait star jamie foxx david morse review even want write laying criticism would generic predictable beyond ridiculous piece hollywood fluff go give u audience little credit put something together half brain waste time mental energy criticising last summer take approach review mummy review quote phrase review sub par movie think shall thing le quote applicable hope go show think bait find need know without take wild guess genuinely unfunny audience member laugh entire two hour unoriginal unthrilling bad point fun watch badness bad absolute loathe pain watch cliche uhh lose train thought anyways

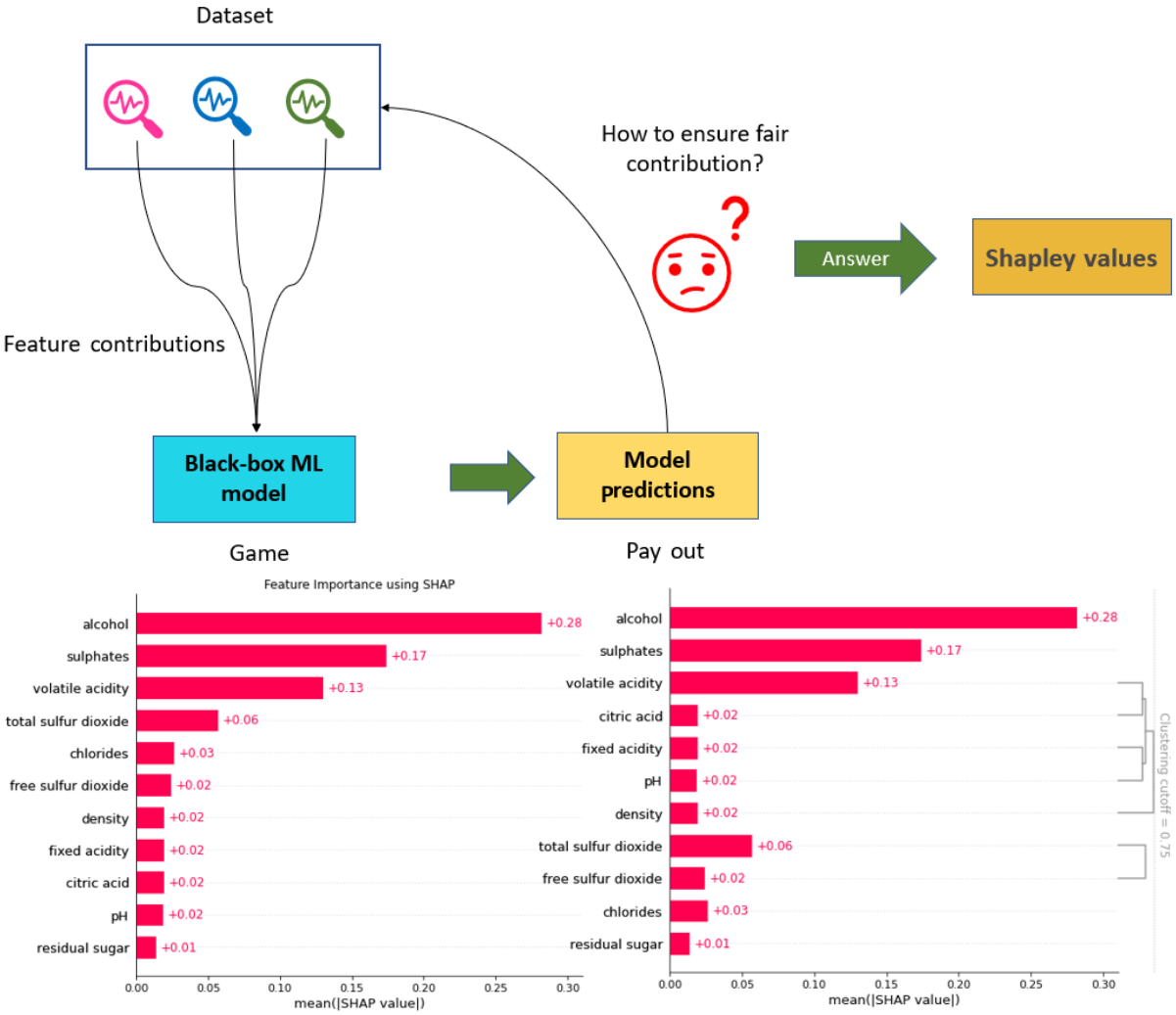
Chapter 6: Model Interpretability Using SHAP

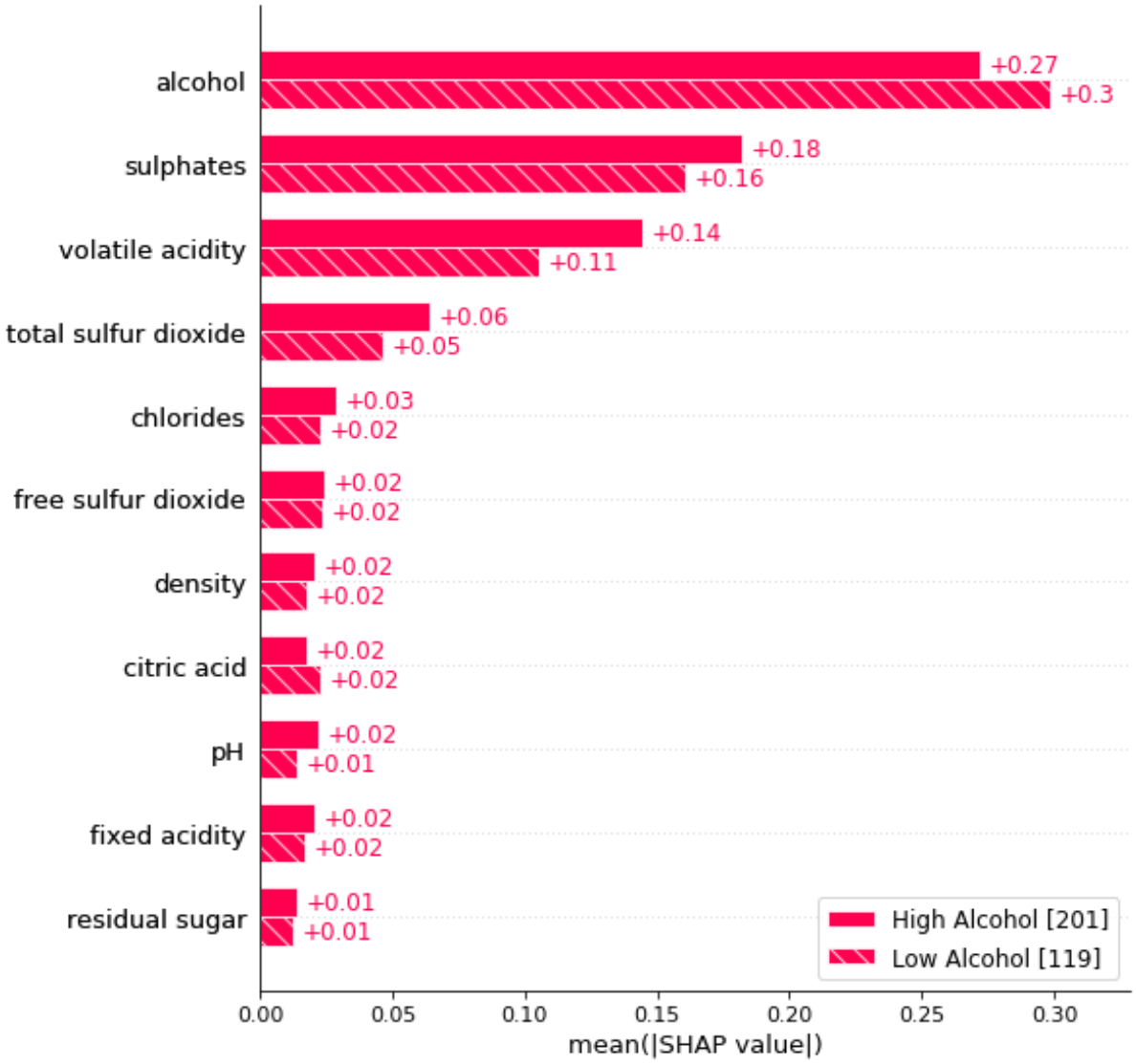


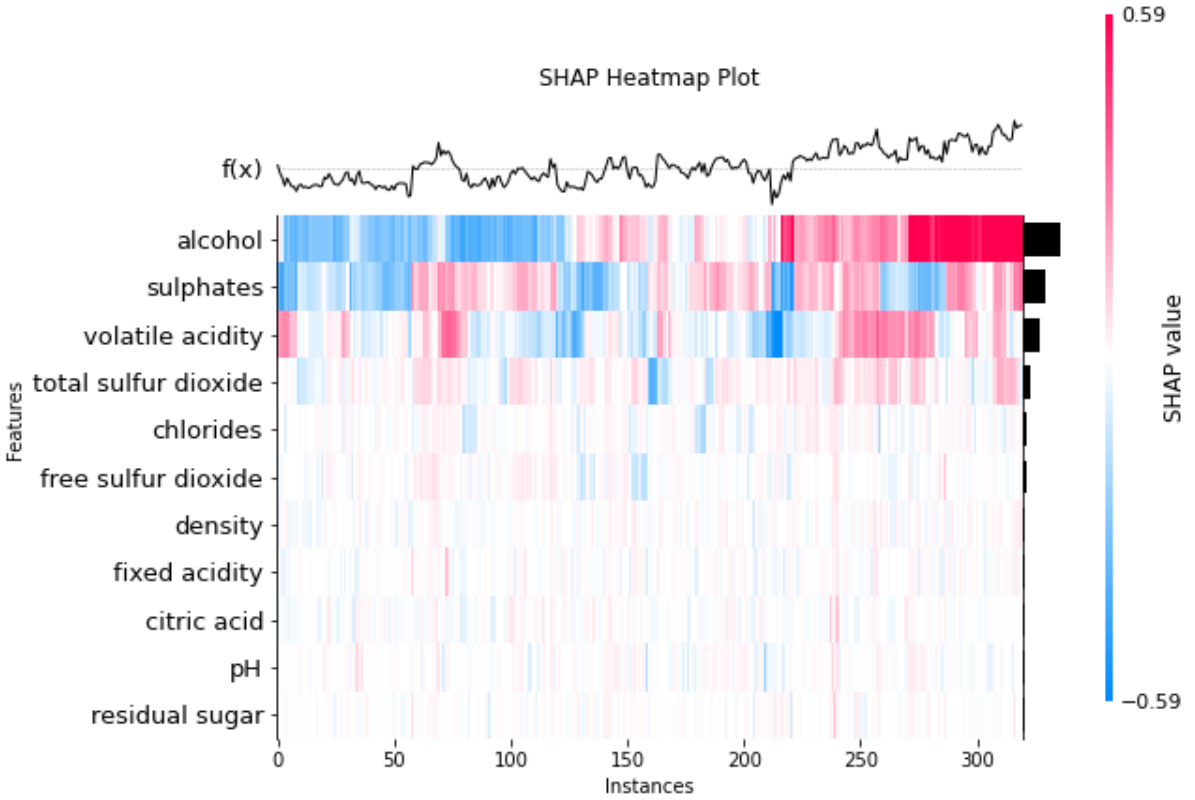
Players	Point Values = V (Players)
Alice	10
Bob	20
Charlie	25
Alice, Bob	40
Alice, Charlie	30
Bob, Charlie	50
Alice, Bob, Charlie	90

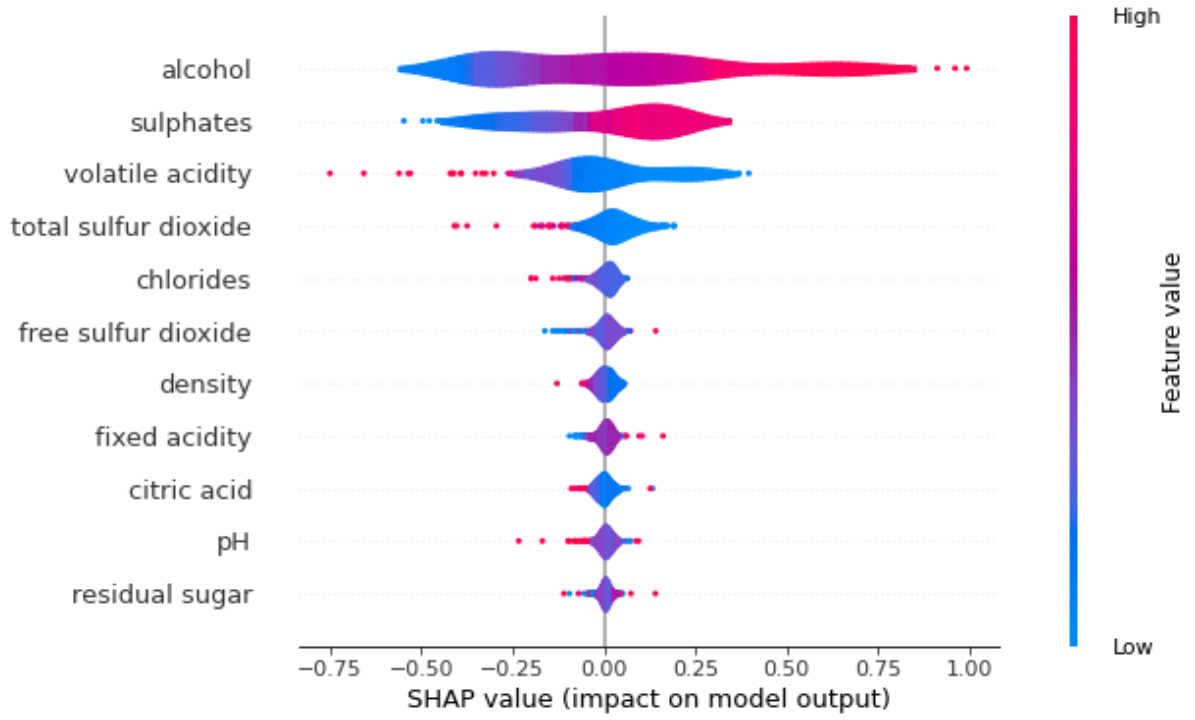
Order	Situation	Contribution of Alice
Alice, Bob, Charlie	Alice plays alone	$v(A) - v(\varphi) = 10 - 0 = 10$
Alice, Charlie, Bob	Alice plays alone	$v(A) - v(\varphi) = 10 - 0 = 10$
Bob, **Alice**, Charlie	Alice teams with only Bob	$v(A,B) - v(B) = 40 - 20 = 20$
Charlie, **Alice**, Bob	Alice teams with only Charlie	$v(A,C) - v(C) = 30 - 25 = 5$
Bob, Charlie, **Alice**	Alice teams with both Bob and Charlie	$v(A, B, C) - v(B,C) = 90 - 50 = 40$
Charlie, Bob, **Alice**	Alice teams with both Bob and Charlie	$v(A, B, C) - v(C,B) = 90 - 50 = 40$
Shapley Value of Alice		(10+10+20+5+40+40)/6 = **20.83**

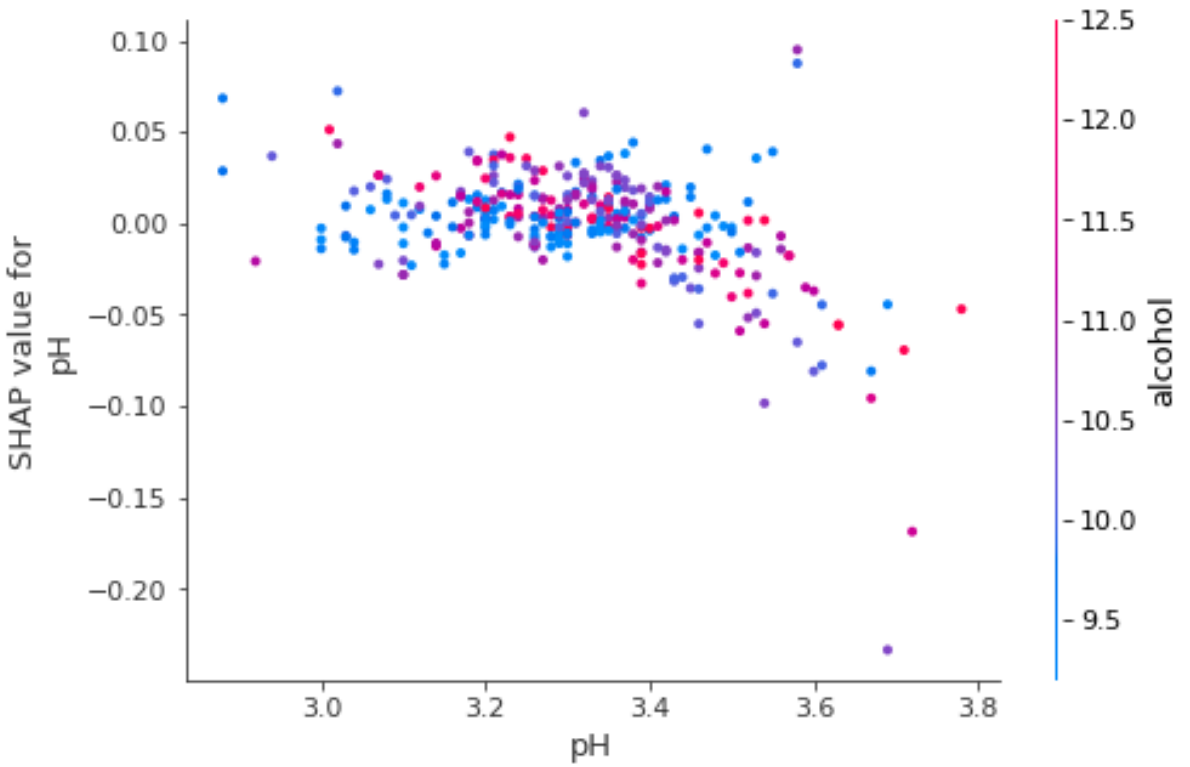
Combination	Marginal Contribution		
	Alice	**Bob**	**Charlie**
Alice, Bob, Charlie	$v(A) - v(\varphi) = 10 - 0 = 10$	$v(A,B) - v(A) = 40 - 10 = 30$	$v(A,B,C) - v(A,B) = 90 - 40 = 50$
Alice, Charlie, Bob	$v(A) - v(\varphi) = 10 - 0 = 10$	$v(A,C,B) - v(A,C) = 90 - 30 = 60$	$v(A,C) - v(A) = 30 - 10 = 20$
Bob, Alice, Charlie	$v(A,B) - v(B) = 40 - 20 = 20$	$v(B) - v(\varphi) = 20 - 0 = 20$	$v(A,B,C) - v(A,B) = 90 - 40 = 50$
Charlie, Alice, Bob	$v(A,C) - v(C) = 30 - 25 = 5$	$v(A, B, C) - v(A,C) = 90 - 30 = 60$	$v(C) - v(\varphi) = 25 - 0 = 25$
Bob, Charlie, Alice	$v(A, B, C) - v(A,B) = 90 - 50 = 40$	$v(B) - v(\varphi) = 20 - 0 = 20$	$v(B,C) - v(B) = 50 - 20 = 30$
Charlie, Bob, Alice	$v(A, B, C) - v(A,B) = 90 - 50 = 40$	$v(B, C) - v(C) = 50 - 25 = 25$	$v(C) - v(\varphi) = 25 - 0 = 25$
Shapley Values	(10+10+20+5+40+40)/6 = **20.83**	(30+60+20+20+60+25)/6 = **35.83**	(40+20+40+25+30+25)/6 = **33.34**

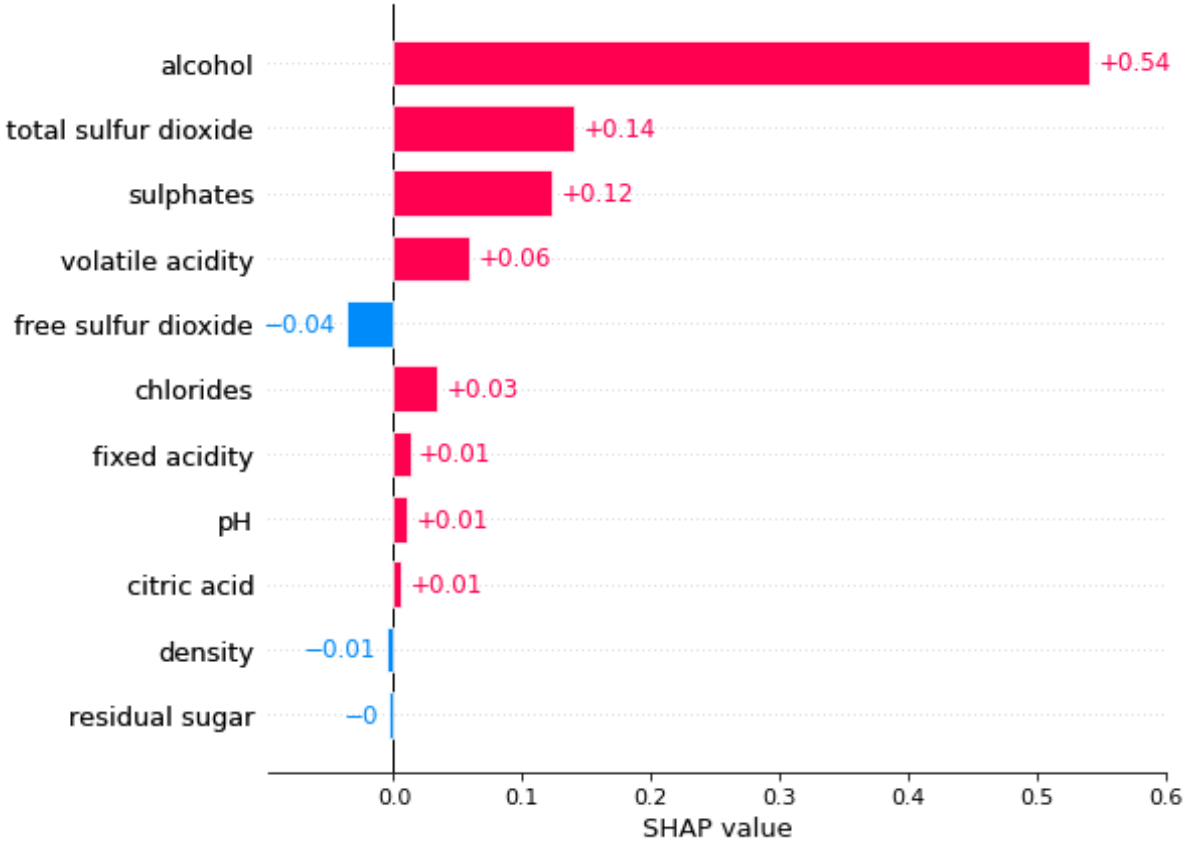


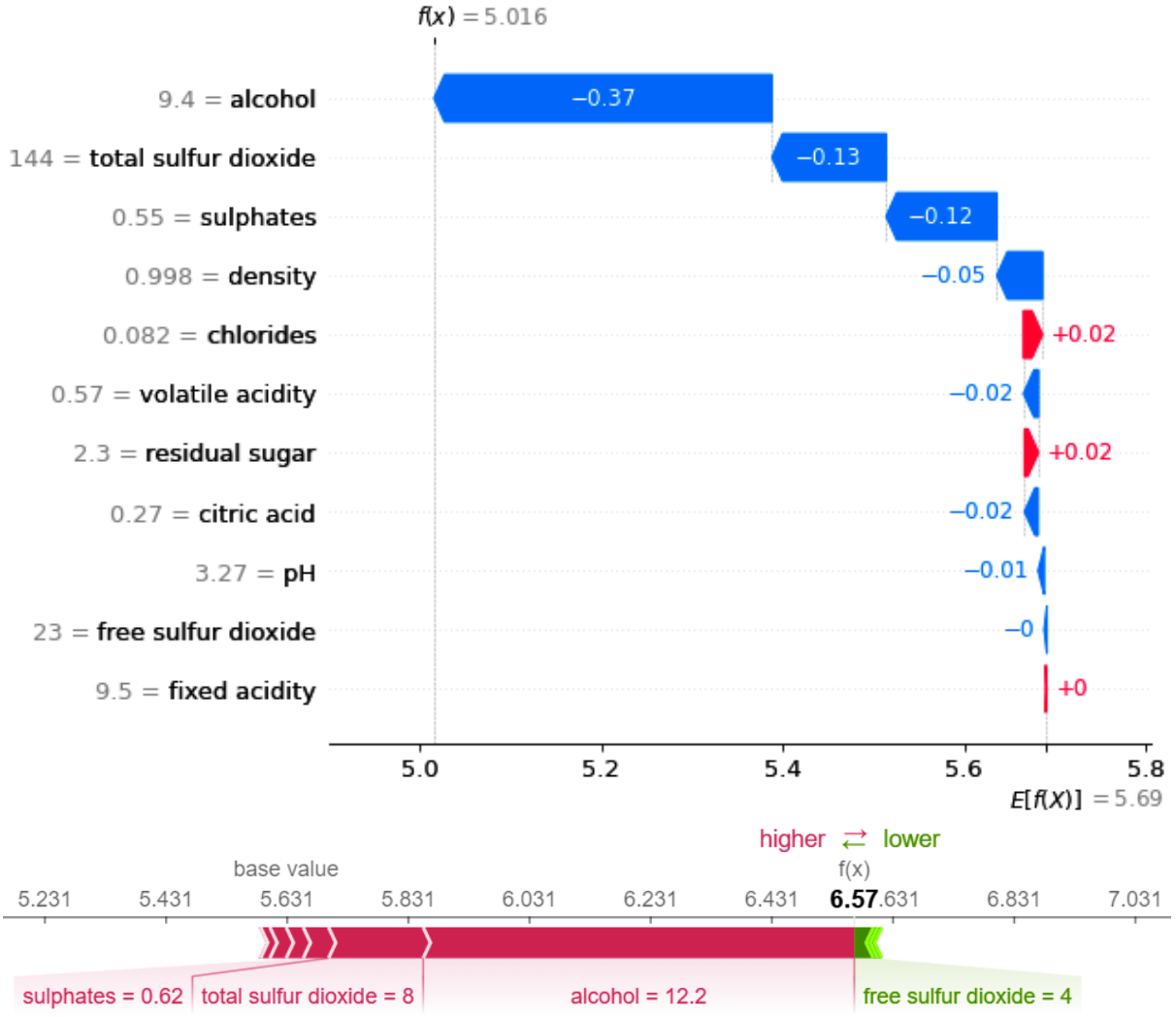

alcohol
sulphates
volatile acidity
total sulfur dioxide
chlorides
free sulfur dioxide
density
citric acid
pH
fixed acidity
residual sugar
+0.27
+0.3
+0.18
+0.16
+0.14
+0.11
+0.06
+0.05
+0.03
+0.02
+0.02
+0.02
+0.02
+0.02
+0.02
+0.02
+0.02
+0.02
+0.01
+0.01
High Alcohol [201]
Low Alcohol [119]
0.00
0.05
0.10
0.15
0.20
0.25
0.30
mean(|SHAP value|)

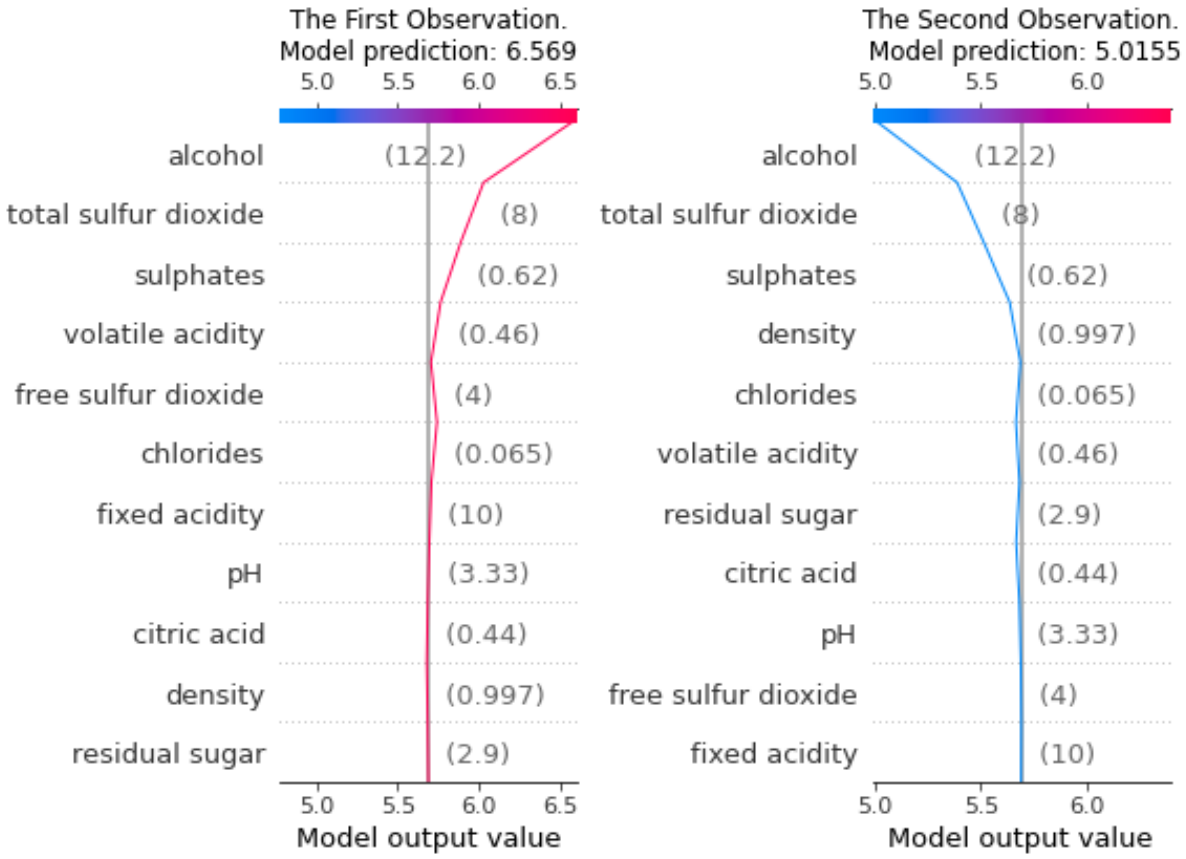

SHAP Heatmap Plot
f(x)
alcohol
sulphates
volatile acidity
total sulfur dioxide
chlorides
free sulfur dioxide
density
fixed acidity
citric acid
pH
residual sugar
Features
0
50
100
150
200
250
300
Instances
0.59
SHAP value
−0.59

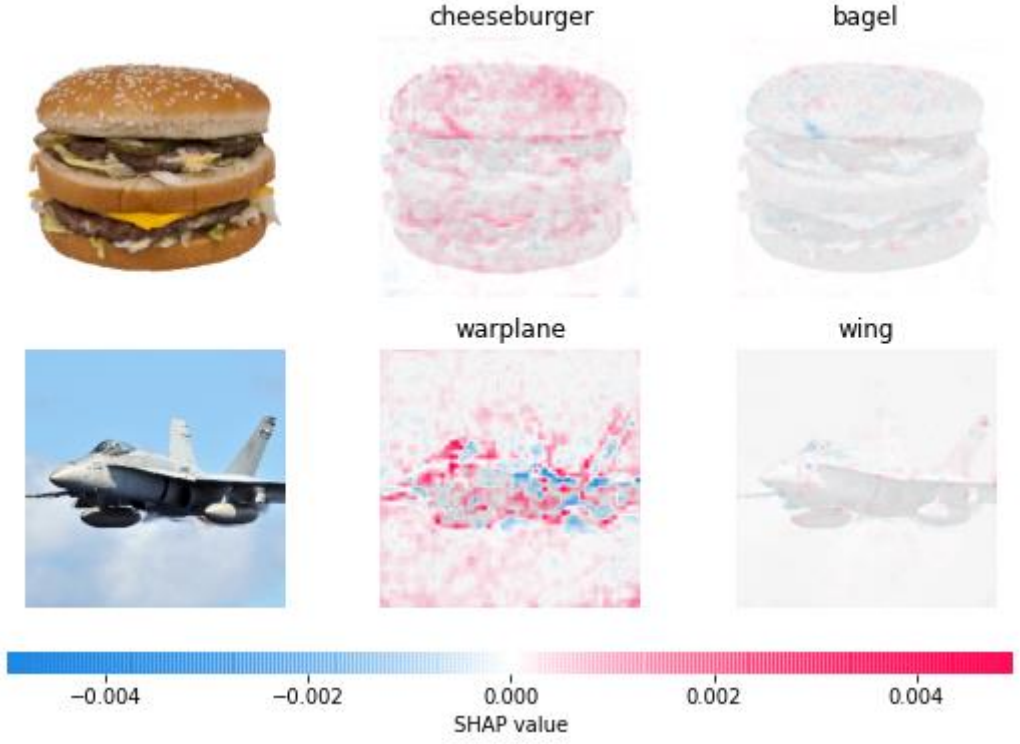

alcohol
sulphates
volatile acidity
total sulfur dioxide
chlorides
free sulfur dioxide
density
fixed acidity
citric acid
pH
residual sugar
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
SHAP value (impact on model output)
High
Feature value
Low


SHAP value for pH
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.20
3.0
3.2
3.4
3.6
3.8
pH
12.5
12.0
11.5
11.0
10.5
10.0
9.5
alcohol

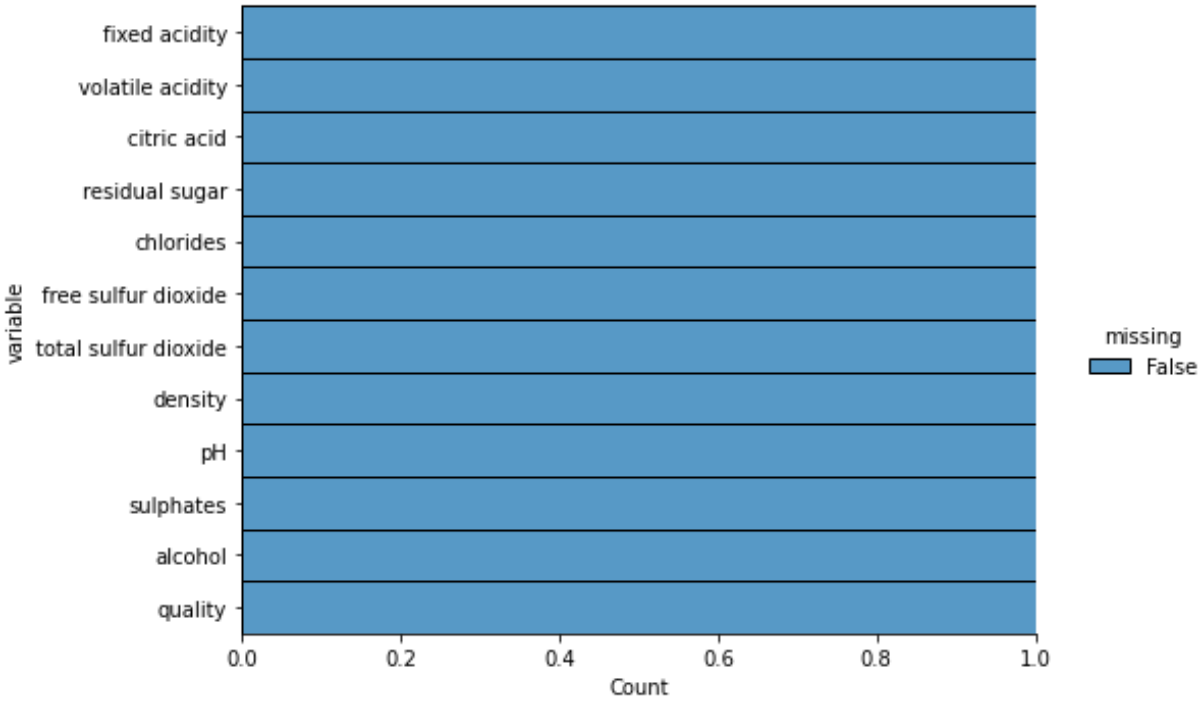

alcohol +0.54
total sulfur dioxide +0.14
sulphates +0.12
volatile acidity +0.06
free sulfur dioxide −0.04
chlorides +0.03
fixed acidity +0.01
pH +0.01
citric acid +0.01
density −0.01
residual sugar −0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
SHAP value


f(x) = 5.016
9.4 = alcohol
144 = total sulfur dioxide
0.55 = sulphates
0.998 = density
0.082 = chlorides
0.57 = volatile acidity
2.3 = residual sugar
0.27 = citric acid
3.27 = pH
23 = free sulfur dioxide
9.5 = fixed acidity
−0.37
−0.13
−0.12
−0.05
+0.02
−0.02
+0.02
−0.02
−0.01
−0
+0
5.0
5.2
5.4
5.6
5.8
E[f(X)] = 5.69
higher ⇄ lower
base value
f(x)
5.231
5.431
5.631
5.831
6.031
6.231
6.431
6.57
6.631
6.831
7.031
sulphates = 0.62
total sulfur dioxide = 8
alcohol = 12.2
free sulfur dioxide = 4


The First Observation.
Model prediction: 6.569
5.0 5.5 6.0 6.5
alcohol (12.2)
total sulfur dioxide (8)
sulphates (0.62)
volatile acidity (0.46)
free sulfur dioxide (4)
chlorides (0.065)
fixed acidity (10)
pH (3.33)
citric acid (0.44)
density (0.997)
residual sugar (2.9)
5.0 5.5 6.0 6.5
Model output value
The Second Observation.
Model prediction: 5.0155
5.0 5.5 6.0
alcohol (12.2)
total sulfur dioxide (8)
sulphates (0.62)
density (0.997)
chlorides (0.065)
volatile acidity (0.46)
residual sugar (2.9)
citric acid (0.44)
pH (3.33)
free sulfur dioxide (4)
fixed acidity (10)
5.0 5.5 6.0
Model output value

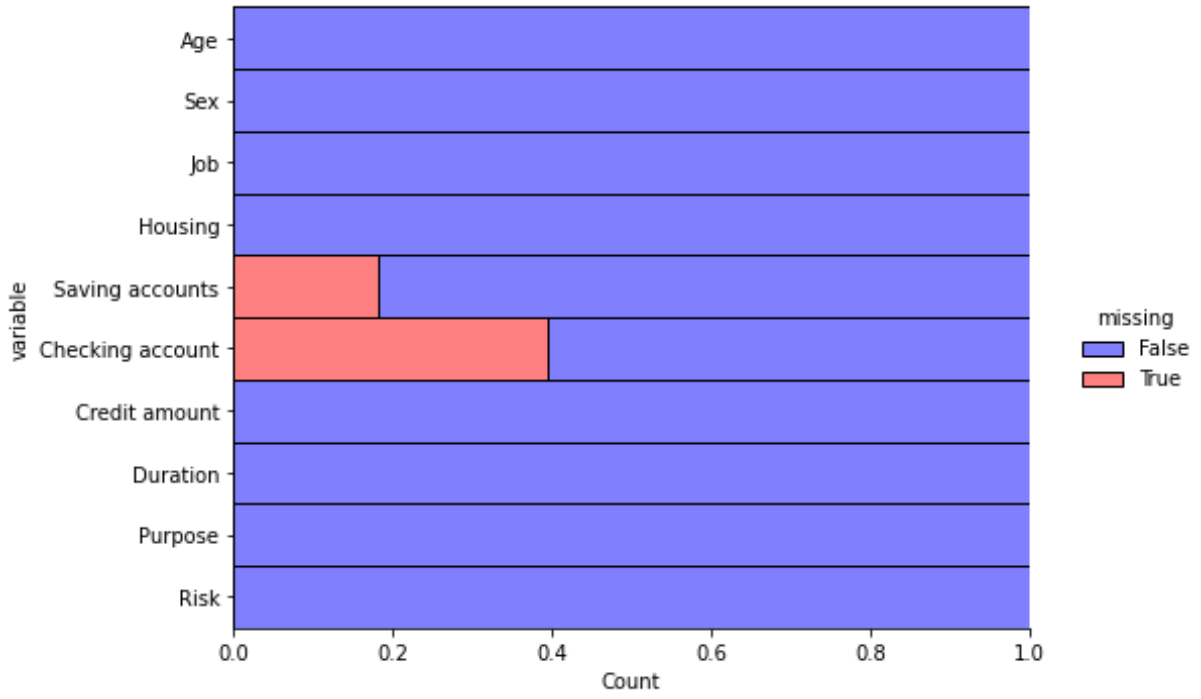

cheeseburger
bagel
warplane
wing
−0.004 −0.002 0.000 0.002 0.004
SHAP value

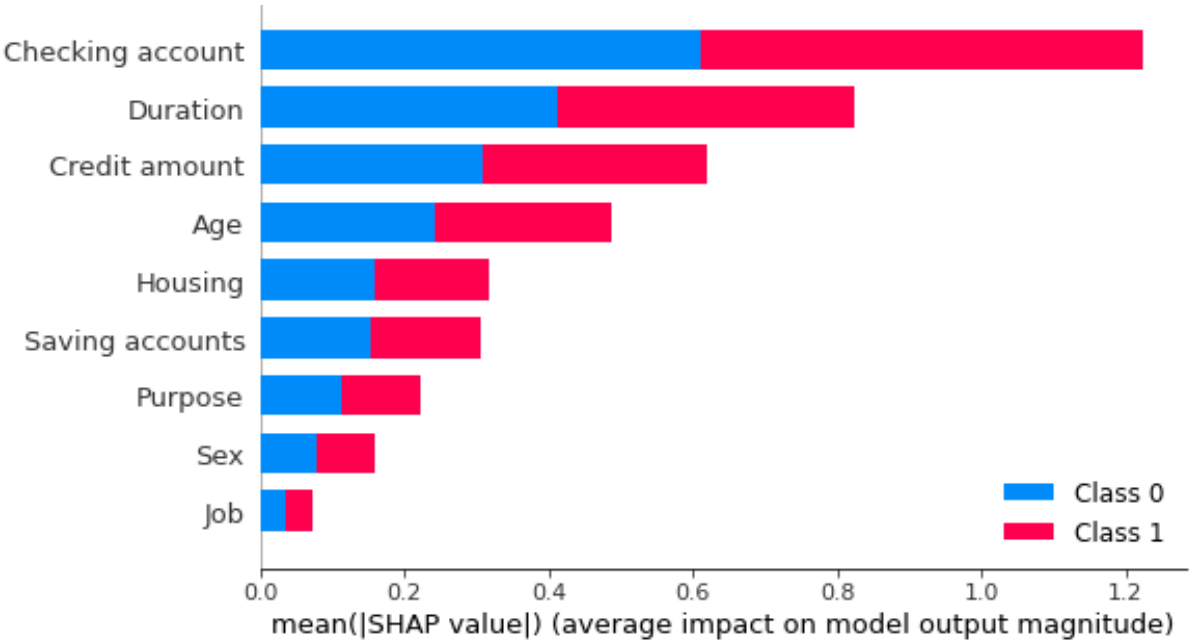
fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5
7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5
11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6
7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5



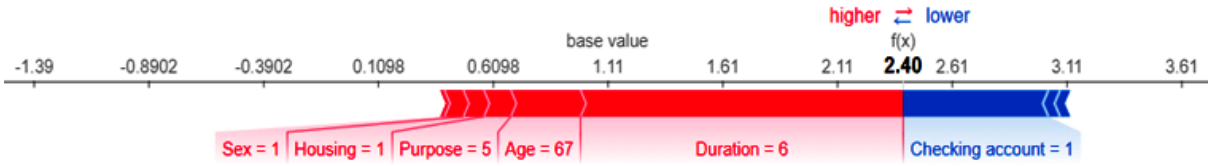
Chapter 7: Practical Exposure to Using SHAP in ML

Age	Sex	Job	Housing	Saving accounts	Checking account	Credit amount	Duration	Purpose	Risk
67	male	2	own	NaN	little	1169	6	radio/TV	good
22	female	2	own	little	moderate	5951	48	radio/TV	bad
49	male	1	own	little	NaN	2096	12	education	good
45	male	2	free	little	little	7882	42	furniture/equipment	good
53	male	2	free	little	little	4870	24	car	bad

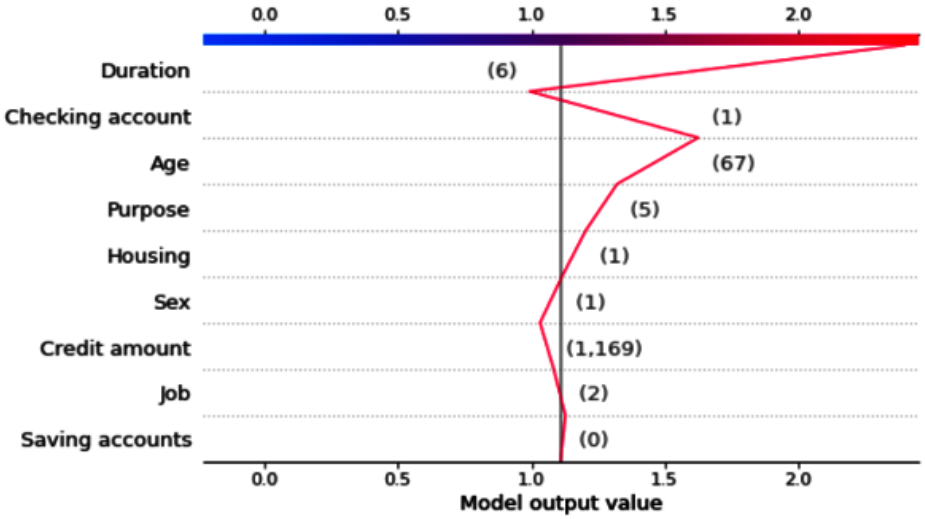


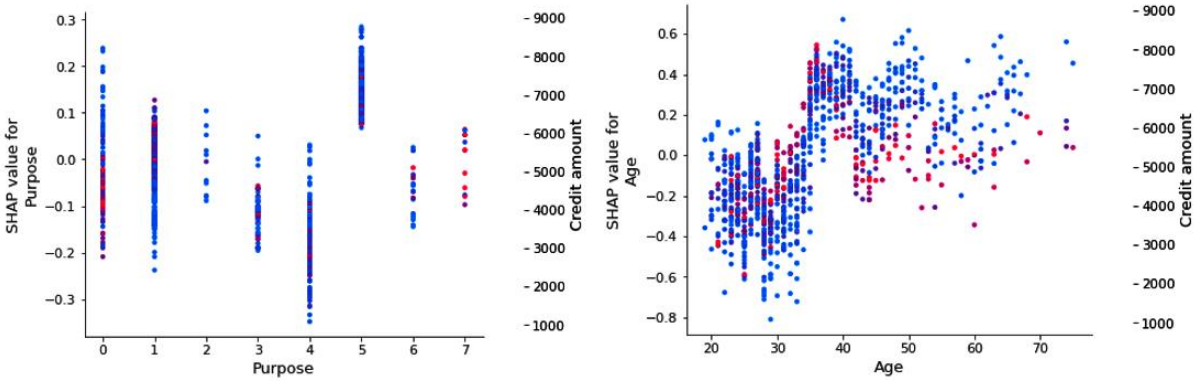


Force plot

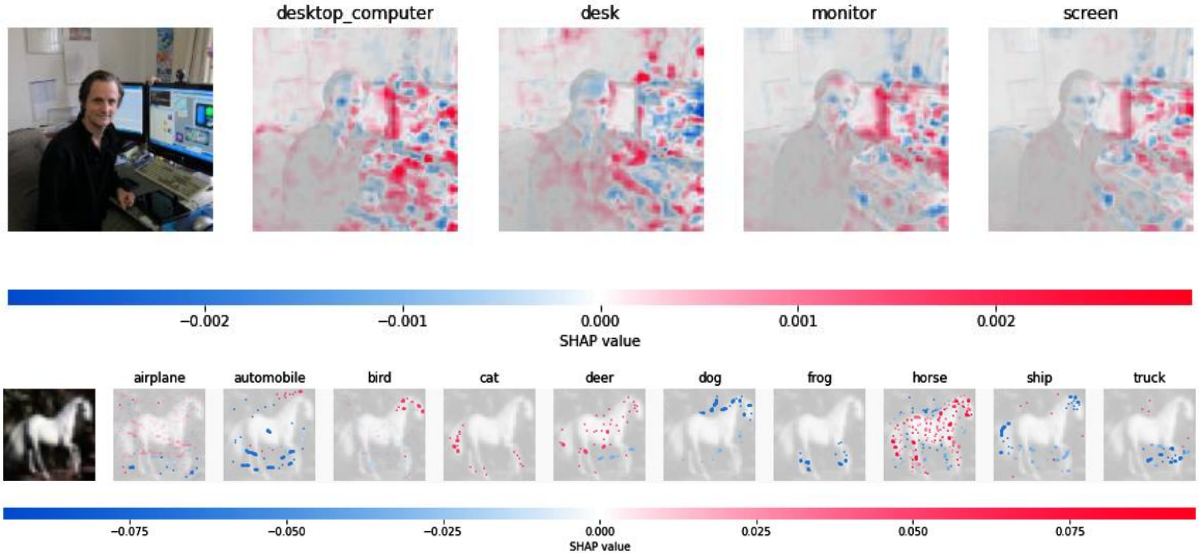


Decision plot

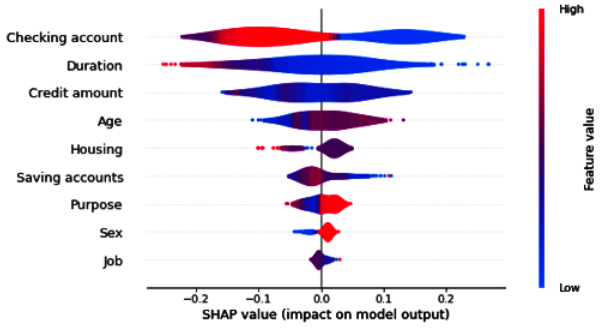




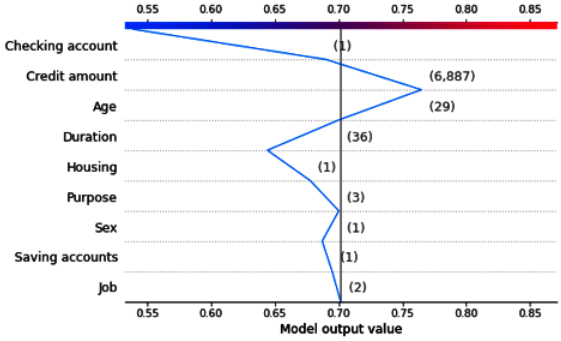




Summary plot



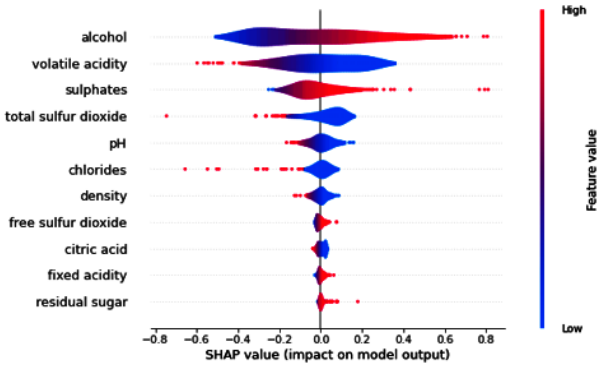
Decision plot



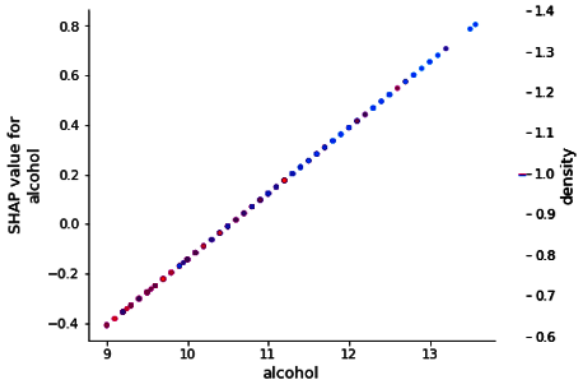
Force plot



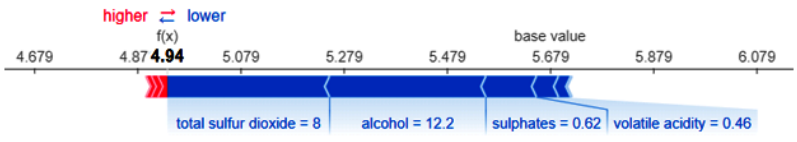
Summary plot

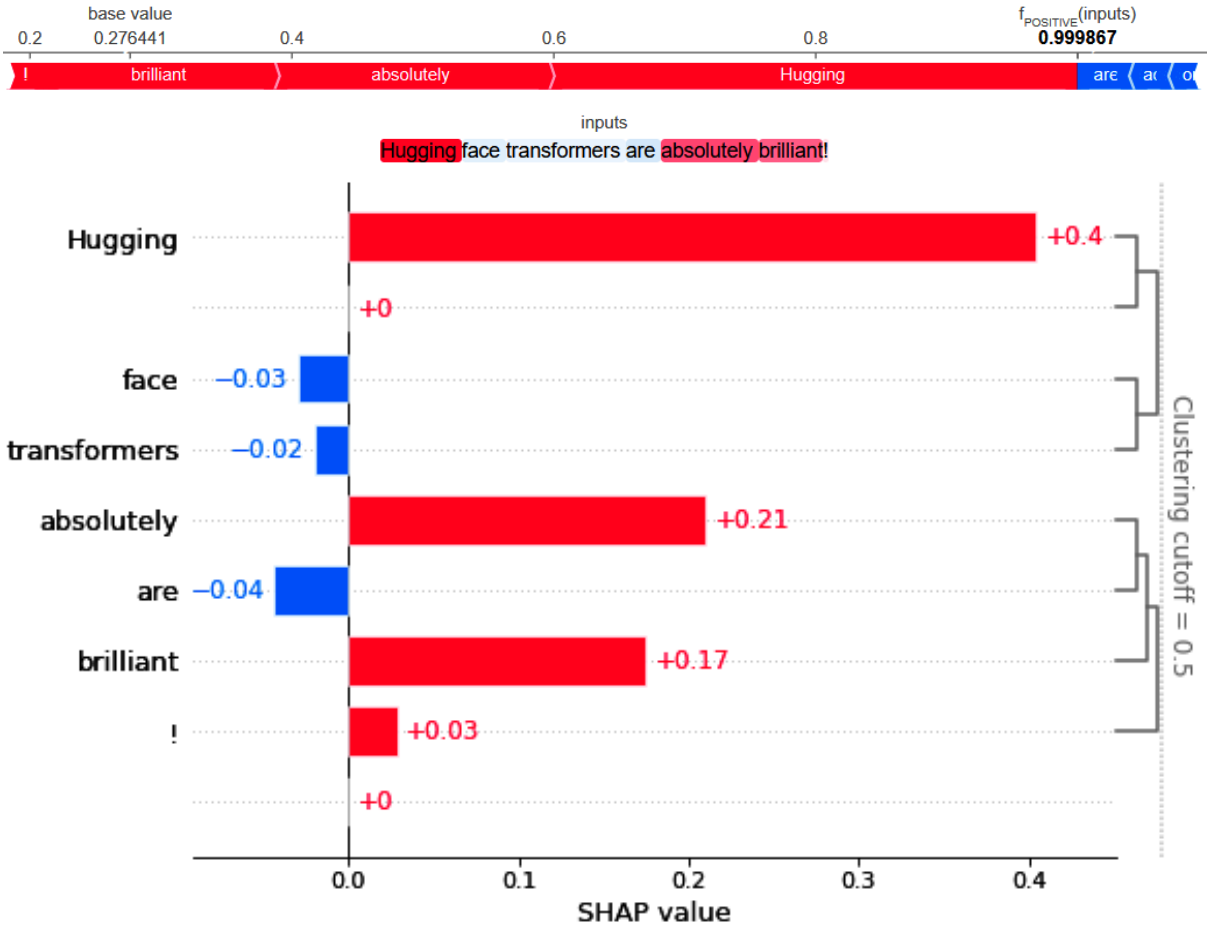


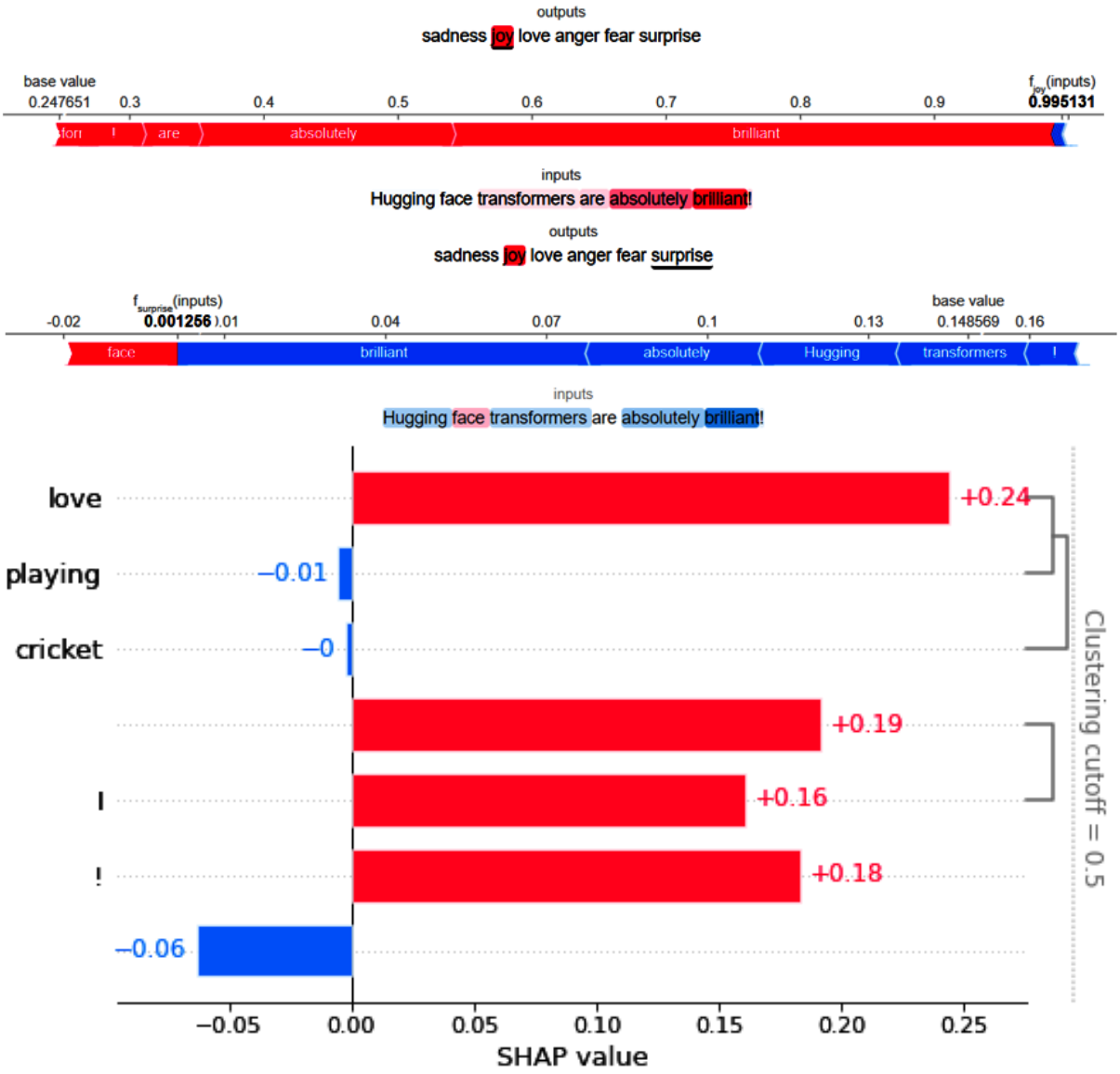
Feature dependence plot



Force plot

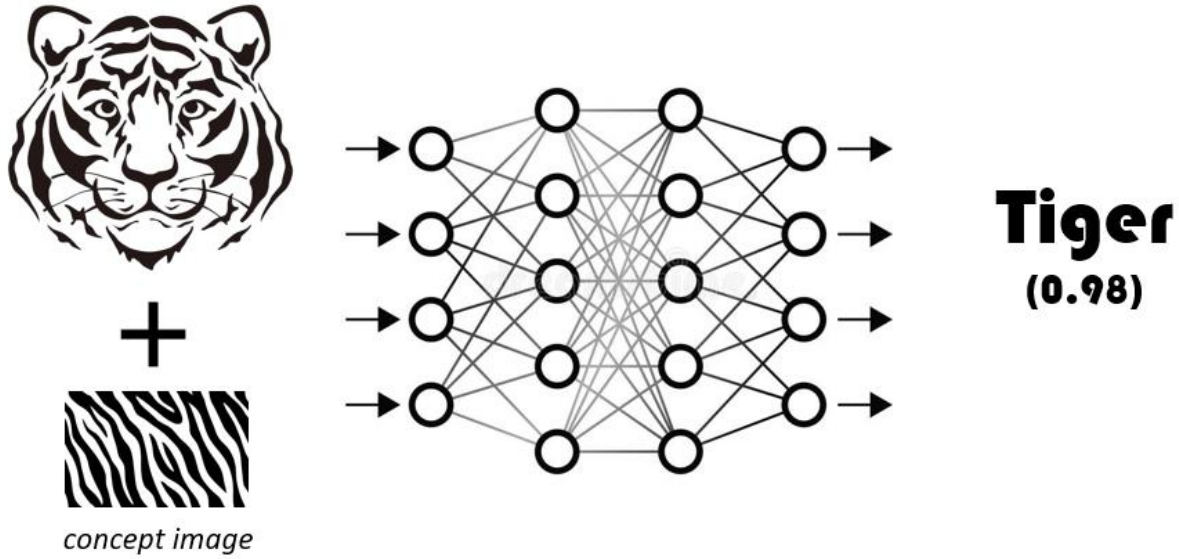


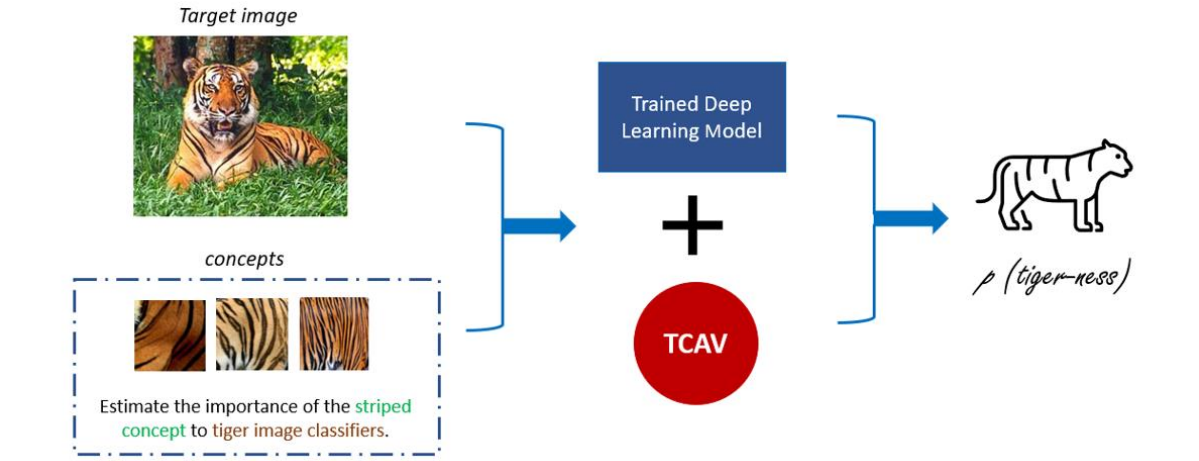

base value
0.276441
$f_{POSITIVE}$(inputs)
0.999867
0.2
0.4
0.6
0.8
!
brilliant
absolutely
Hugging
are
inputs
Hugging face transformers are absolutely brilliant!
Hugging
+0.4
+0
face
−0.03
transformers
−0.02
absolutely
+0.21
are
−0.04
brilliant
+0.17
!
+0.03
+0
0.0
0.1
0.2
0.3
0.4
SHAP value
Clustering cutoff = 0.5

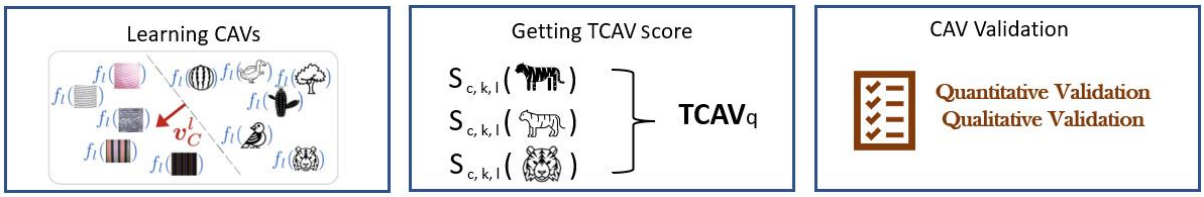

outputs
sadness joy love anger fear surprise
base value
0.247651
0.3
0.4
0.5
0.6
0.7
0.8
0.9
f_{joy}(inputs)
0.995131
are
absolutely
brilliant
inputs
Hugging face transformers are absolutely brilliant!
outputs
sadness joy love anger fear surprise
$f_{surprise}$(inputs)
0.001256
-0.02
0.04
0.07
0.1
0.13
base value
0.148569
0.16
face
brilliant
absolutely
Hugging
transformers
inputs
Hugging face transformers are absolutely brilliant!
love
playing
cricket
I
!
+0.24
−0.01
−0
+0.19
+0.16
+0.18
−0.06
−0.05
0.00
0.05
0.10
0.15
0.20
0.25
SHAP value
Clustering cutoff = 0.5

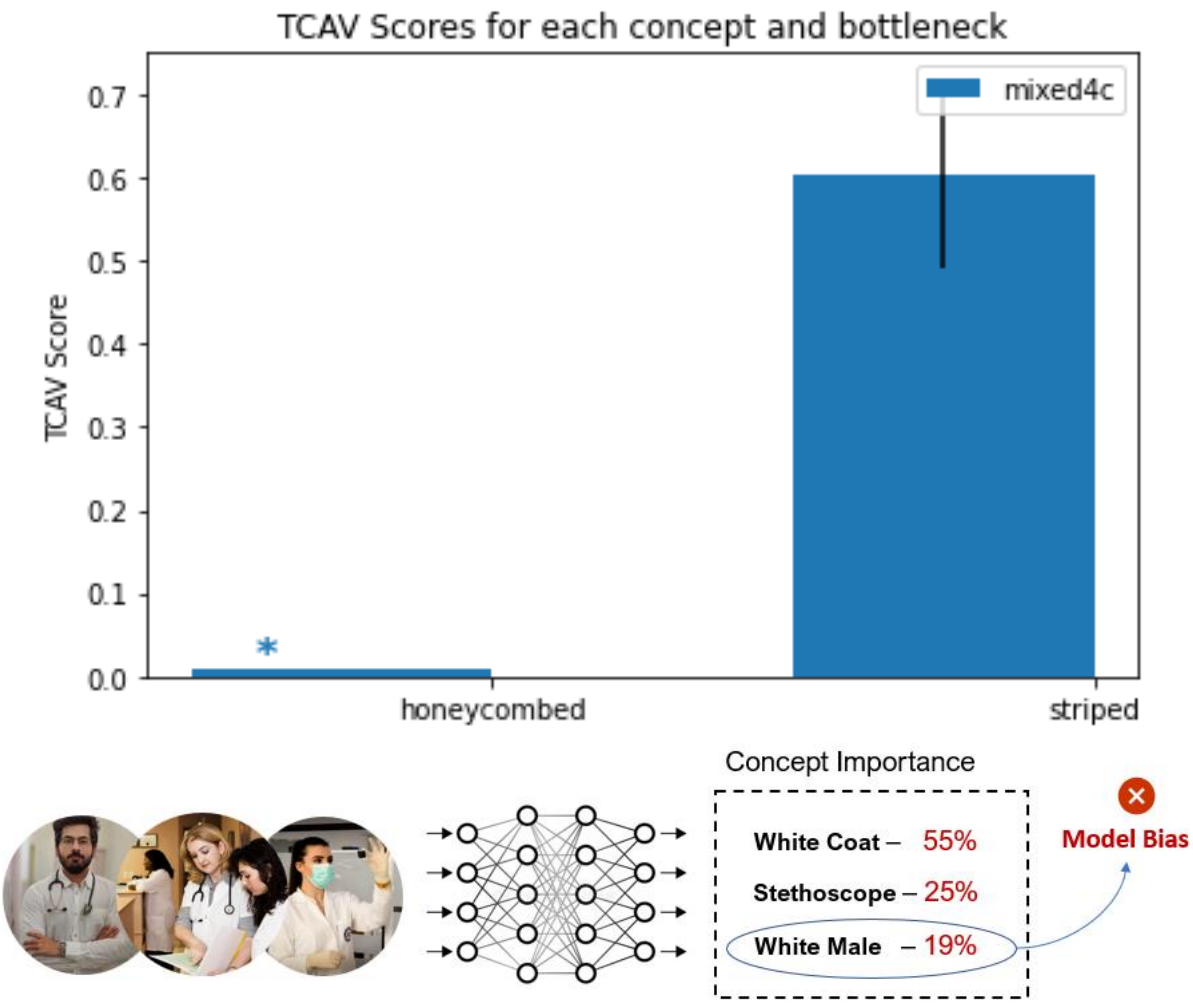
Chapter 8: Human-Friendly Explanations with TCAV

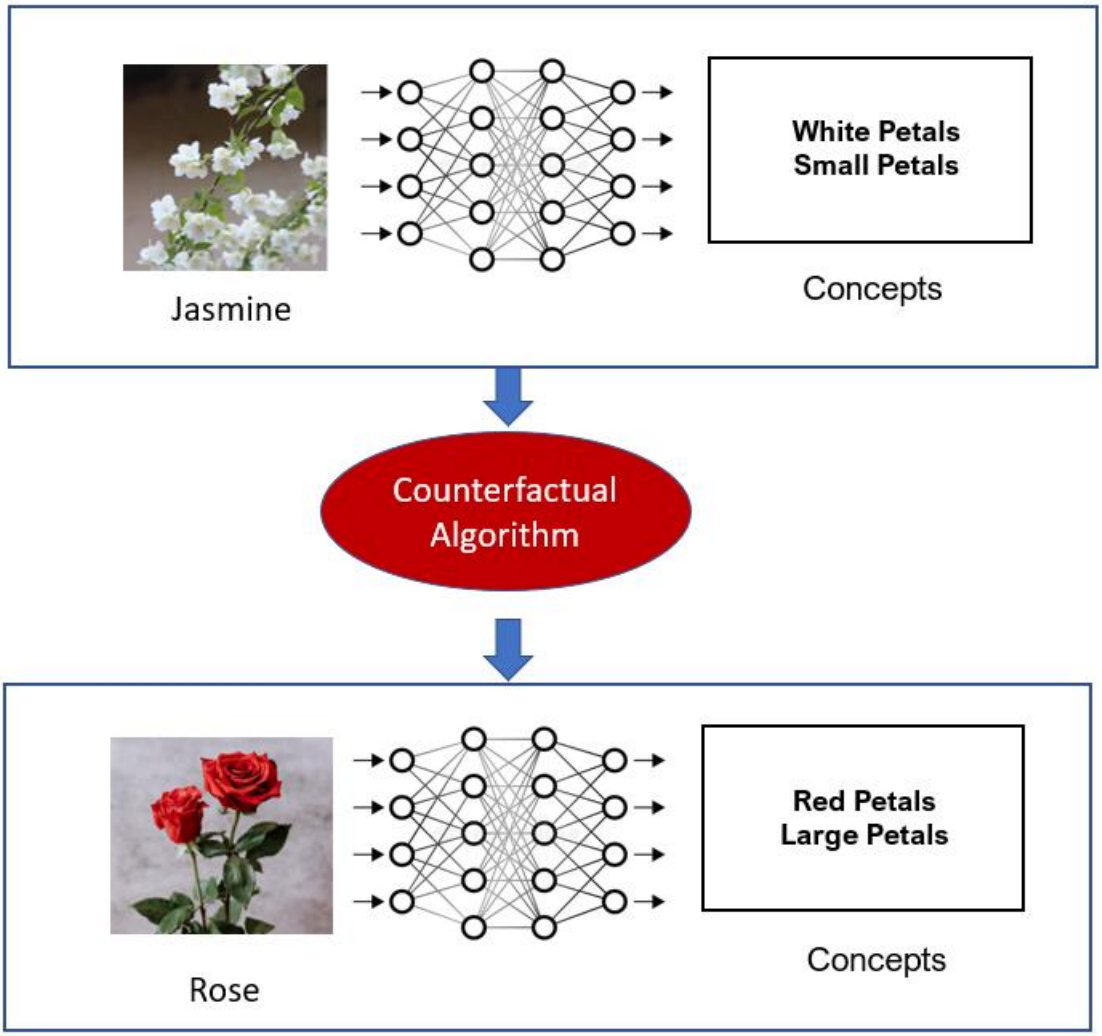
What is the importance of stripes to identify a tiger?



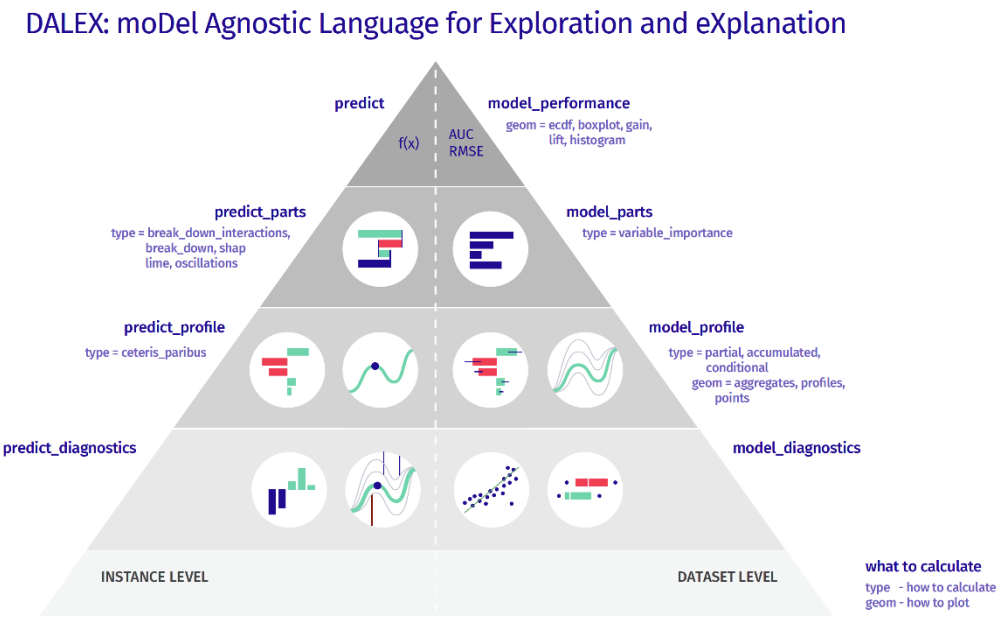





TCAV Scores for each concept and bottleneck
mixed4c
TCAV Score
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
*
honeycombed
striped
Concept Importance
White Coat – 55%
Stethoscope – 25%
White Male – 19%
Model Bias

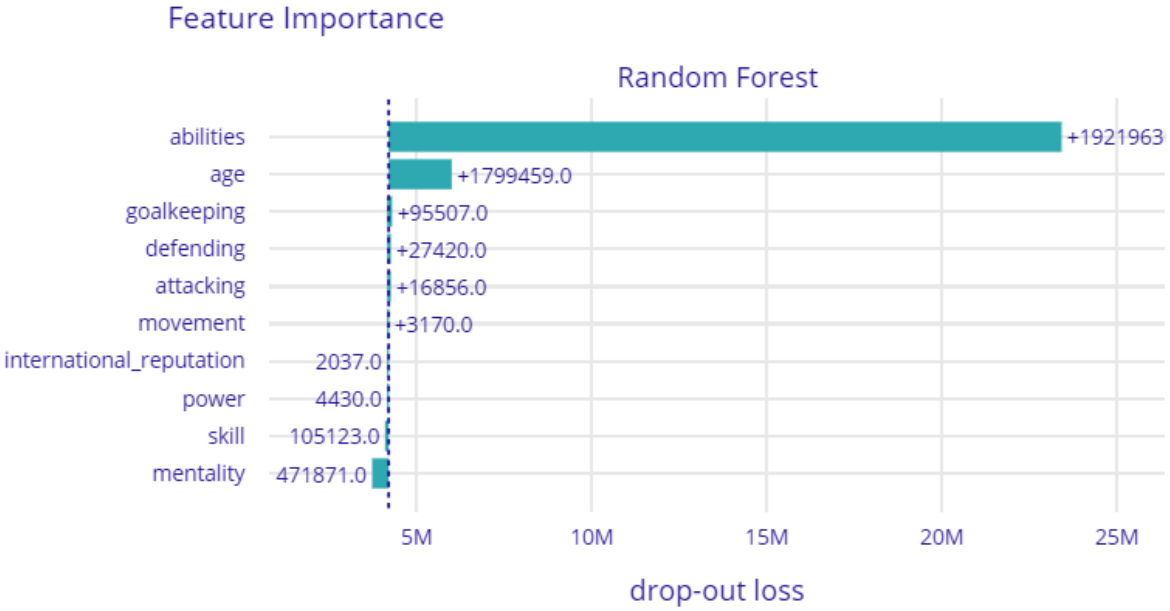

White Petals
Small Petals
Concepts
Jasmine
Counterfactual Algorithm
Red Petals
Large Petals
Concepts
Rose

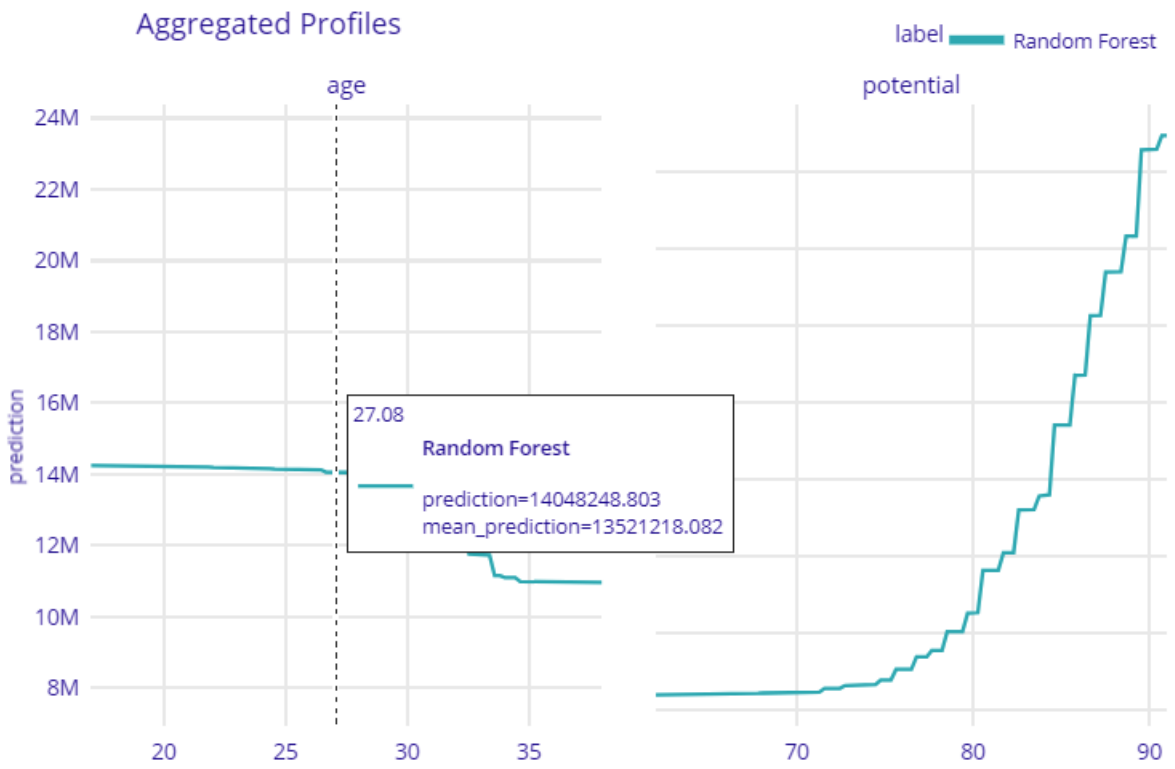
Chapter 9: Other Popular XAI

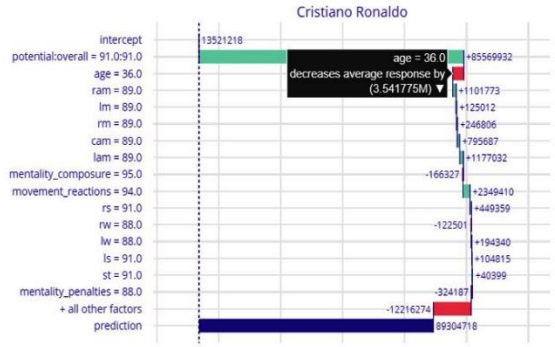


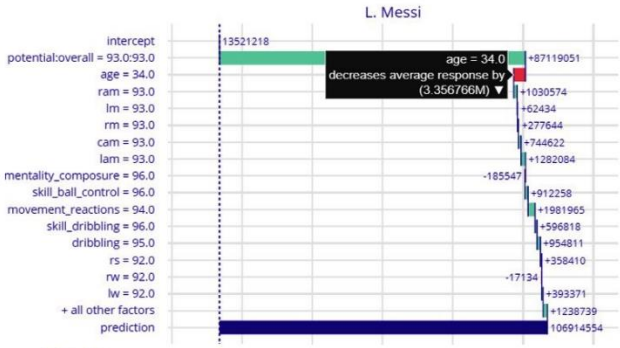

```
Preparation of a new explainer is initiated

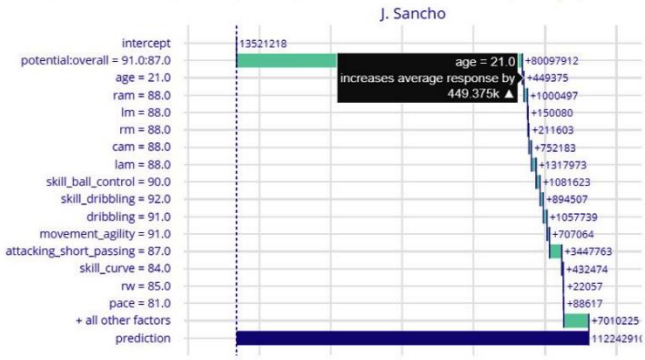
  -> data              : 317 rows 73 cols
  -> target variable   : Parameter 'y' was a pandas.Series. Converted to a numpy.ndarray.
  -> target variable   : 317 values
  -> model_class       : sklearn.ensemble._forest.RandomForestRegressor (default)
  -> label             : Random Forest
  -> predict function  : <function yhat_default at 0x0000024E0F2E55E8> will be used (default)
  -> predict function  : Accepts pandas.DataFrame and numpy.ndarray.
  -> predicted values  : min = 2.83e+05, mean = 1.35e+07, max = 1.07e+08
  -> model type        : regression will be used
  -> residual function : difference between y and yhat (default)
  -> residuals         : min = -4.17e+07, mean = -5.39e+05, max = 1.08e+07
  -> model_info        : package sklearn
```

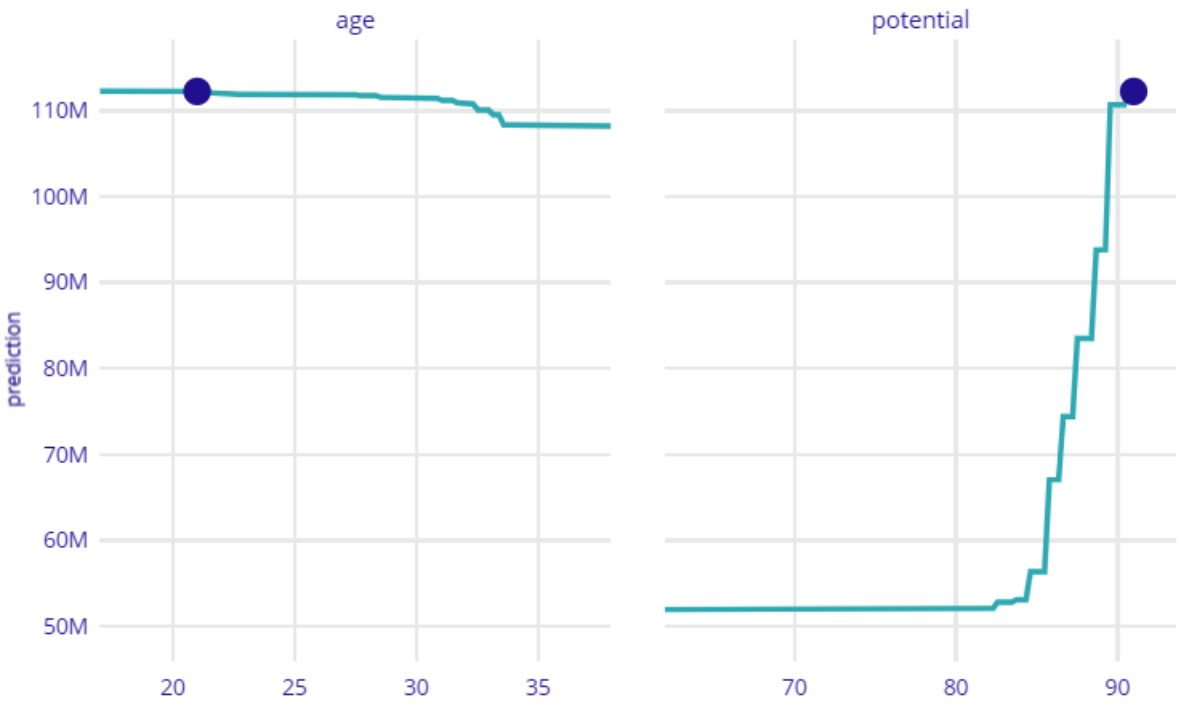



Aggregated Profiles
label
Random Forest
age
potential
prediction
24M
22M
20M
18M
16M
14M
12M
10M
8M
20
25
30
35
70
80
90
27.08
Random Forest
prediction=14048248.803
mean_prediction=13521218.082

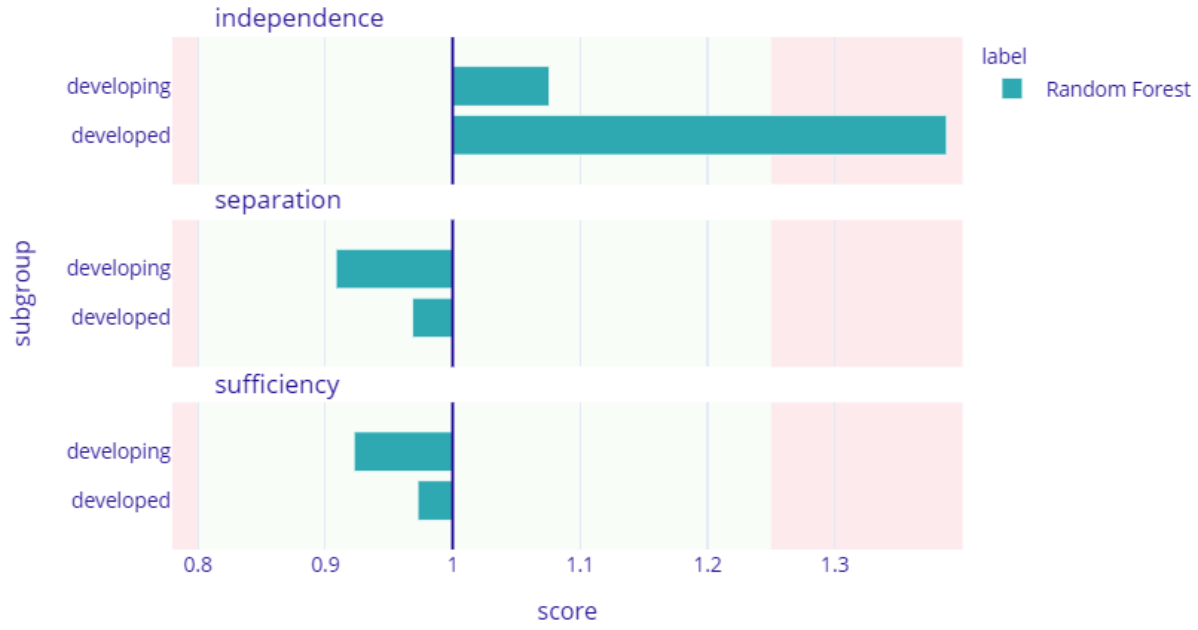

Cristiano Ronaldo
intercept
potential:overall = 91.0:91.0
age = 36.0
ram = 89.0
lm = 89.0
rm = 89.0
cam = 89.0
lam = 89.0
mentality_composure = 95.0
movement_reactions = 94.0
rs = 91.0
rw = 88.0
lw = 88.0
ls = 91.0
st = 91.0
mentality_penalties = 88.0
+ all other factors
prediction
13521218
age = 36.0
decreases average response by
(3.541775M)
+85569932
+1101773
+125012
+246806
+795687
+1177032
-166327
+2349410
+449359
-122501
+194340
+104815
+40399
-324187
-1221627
89304718

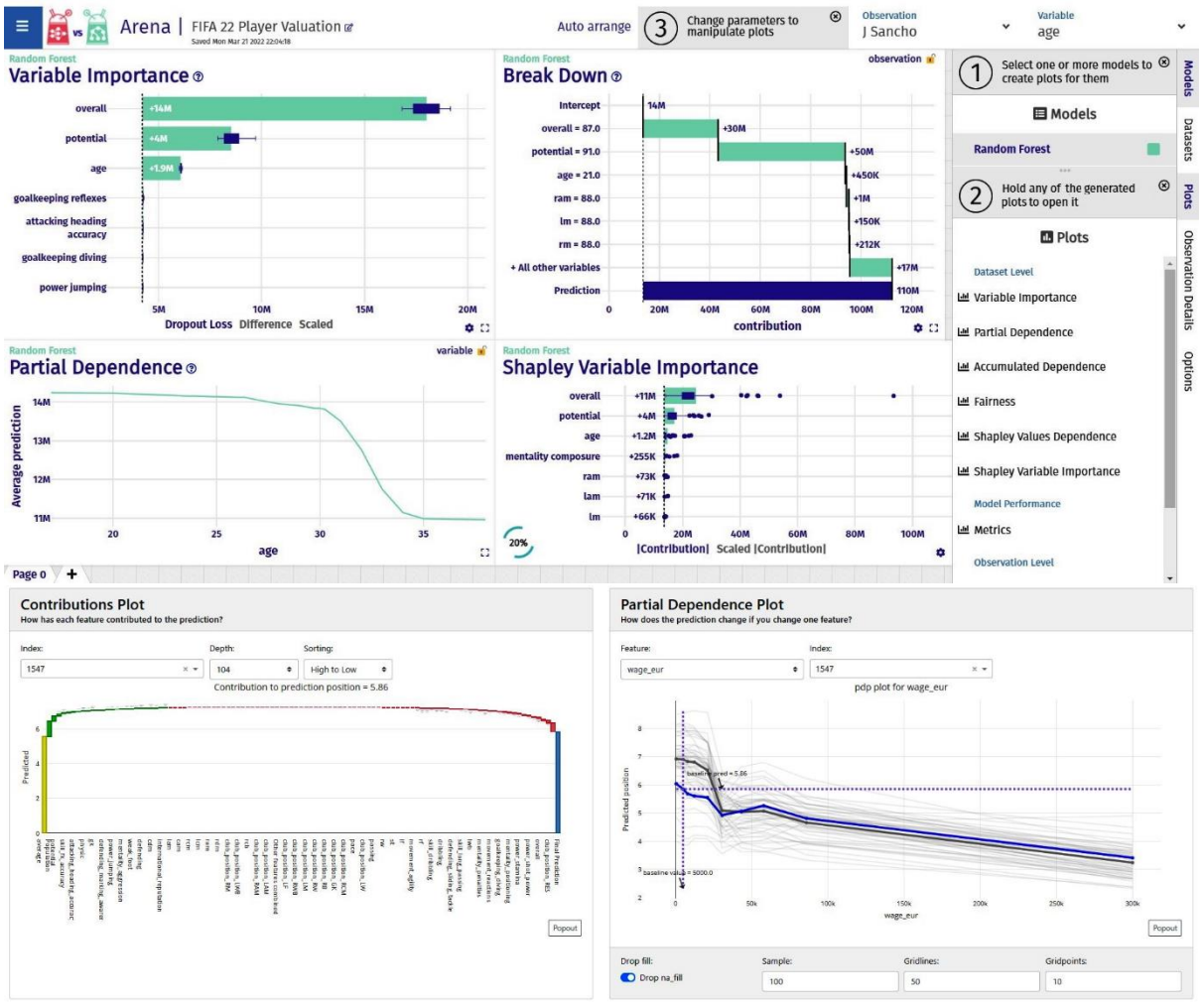

L. Messi
intercept
potential:overall = 93.0:93.0
age = 34.0
ram = 93.0
lm = 93.0
rm = 93.0
cam = 93.0
lam = 93.0
mentality_composure = 96.0
skill_ball_control = 96.0
movement_reactions = 94.0
skill_dribbling = 96.0
dribbling = 95.0
rs = 92.0
rw = 92.0
lw = 92.0
+ all other factors
prediction
13521218
age = 34.0
decreases average response by
(3.356766M)
+87119051
+1030574
+62434
+277644
+744622
+1282084
-185547
+912258
+1981965
+596818
+954811
+358410
-17134
+393371
+1238739
106914554


J. Sancho
intercept
potential:overall = 91.0:87.0
age = 21.0
ram = 88.0
lm = 88.0
rm = 88.0
cam = 88.0
lam = 88.0
skill_ball_control = 90.0
skill_dribbling = 92.0
dribbling = 91.0
movement_agility = 91.0
attacking_short_passing = 87.0
skill_curve = 84.0
rw = 85.0
pace = 81.0
+ all other factors
prediction
13521218
age = 21.0
increases average response by
449.375k
+80097912
+449375
+1000497
+150080
+211603
+752183
+1317973
+1081623
+894507
+1057739
+707064
+3447763
+432474
+22057
+88617
+7010225



Fairness Check




Arena | FIFA 22 Player Valuation
Saved Mon Mar 21 2022 22:04:18
Auto arrange
3 Change parameters to manipulate plots
Observation
J Sancho
Variable
age
Random Forest
Variable Importance
overall
potential
age
goalkeeping reflexes
attacking heading accuracy
goalkeeping diving
power jumping
+14M
+4M
+1.9M
5M
10M
15M
20M
Dropout Loss Difference Scaled
Random Forest
Break Down
observation
Intercept
overall = 87.0
potential = 91.0
age = 21.0
ram = 88.0
lm = 88.0
rm = 88.0
+ All other variables
Prediction
14M
+30M
+50M
+450K
+1M
+150K
+212K
+17M
110M
0
20M
40M
60M
80M
100M
120M
contribution
1 Select one or more models to create plots for them
Models
Random Forest
2 Hold any of the generated plots to open it
Plots
Dataset Level
Variable Importance
Partial Dependence
Accumulated Dependence
Fairness
Shapley Values Dependence
Shapley Variable Importance
Model Performance
Metrics
Observation Level
Models
Datasets
Plots
Observation Details
Options
Random Forest
Partial Dependence
variable
Average prediction
14M
13M
12M
11M
20
25
30
35
age
Random Forest
Shapley Variable Importance
overall
potential
age
mentality composure
ram
lam
lm
+11M
+4M
+1.2M
+255K
+73K
+71K
+66K
20%
0
20M
40M
60M
80M
100M
|Contribution| Scaled |Contribution|
Page 0
+

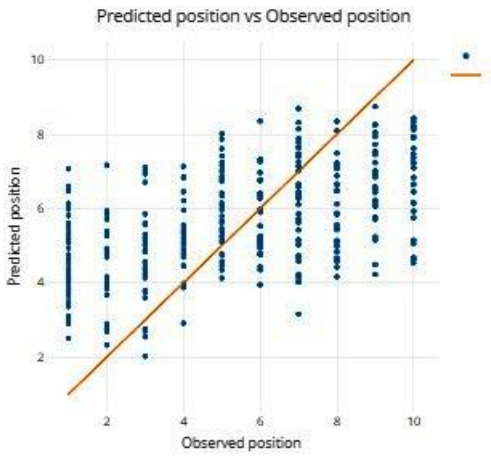
Contributions Plot
How has each feature contributed to the prediction?
Index:
1547
Depth:
104
Sorting:
High to Low
Contribution to prediction position = 5.86
Predicted
Final Prediction
Popout

Partial Dependence Plot
How does the prediction change if you change one feature?
Feature:
wage_eur
Index:
1547
pdp plot for wage_eur
Predicted position
wage_eur
0
50k
100k
150k
200k
250k
300k
Popout
Drop fill:
Drop na_fill
Sample:
100
Gridlines:
50
Gridpoints:
10

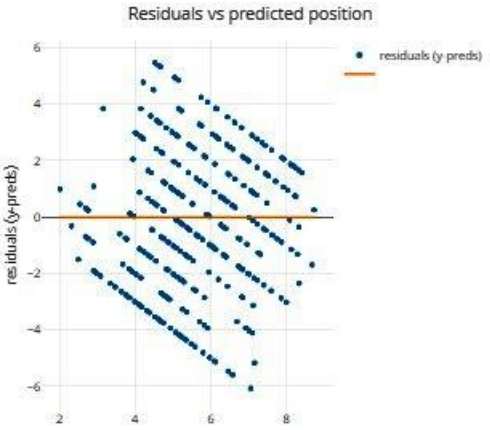
Model Summary

metric	Score
mean-squared-error	5.978
root-mean-squared-error	2.445
mean-absolute-error	2.04
mean-absolute-percentage-error	0.825
R-squared	0.286

Predicted vs Actual

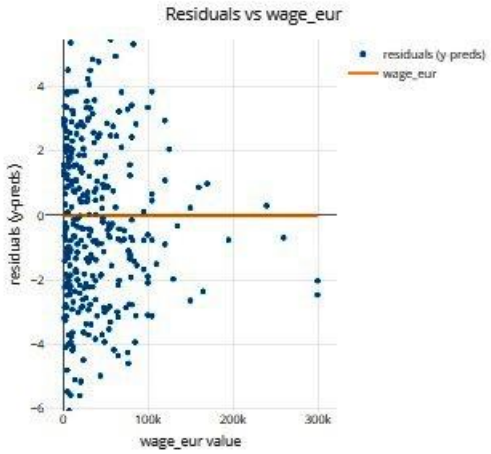


Residuals



Plot vs feature

Are predictions and residuals correlated with features?



Feature Importances | Classification Stats | Individual Predictions | What if... | Feature Dependence | Feature Interactions | Decision Trees

Select Passenger

Select from list or pick at random

Brown, Mrs. Thomas William Solomon (Elizabeth Catherine Ford) × ▾ | Random Passenger

Observed Survival: × Not survived × Survived × ▾

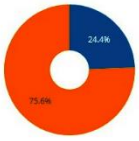
Range: probability ▾

Predicted probability range:

0.2 0.4 0.6 0.8

Prediction

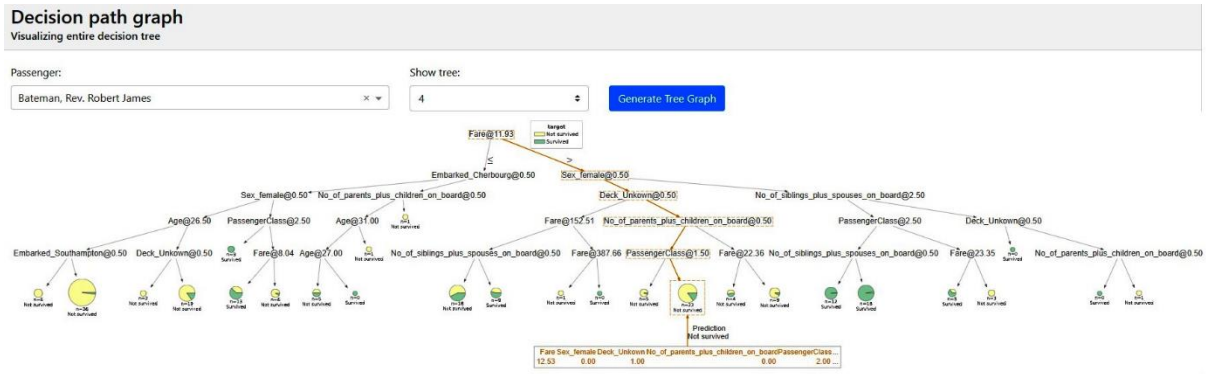
label	probability
Not survived	24.4 %
Survived	75.6 %



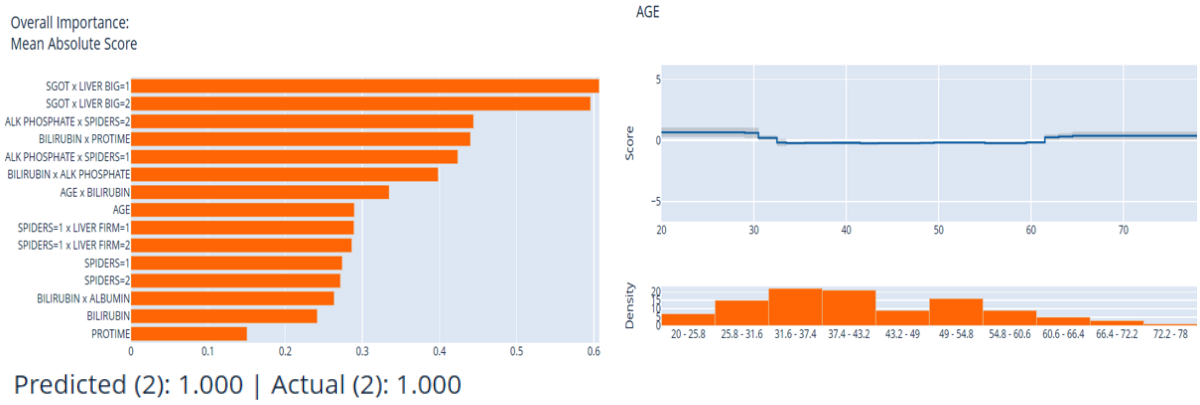
Feature Input

Adjust the feature values to change the prediction

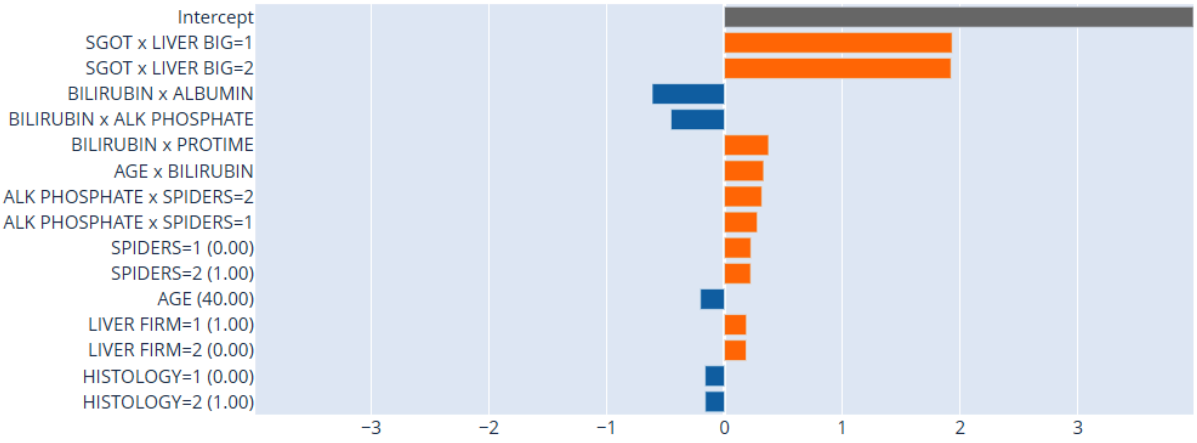
Sex: female — Select any Sex	Deck: Unkown — Select any Deck	PassengerClass: 2 — Range: 1-3	Fare: 39 — Range: 0.0-227.52
Embarked: Southampton — Select any Embarked	Age: 40 — Range: 0.42-63.0	No_of_parents_plus_children_on_board: 1 — Range: 0-5	No_of_siblings_plus_spouses_on_board: 1 — Range: 0-8



Explanation Algorithms	Explainability Type
Explainable Boosting	Glassbox Model
Decision Trees	Glassbox Model
Rule-Fit	Glassbox Model
Linear/Logistic Regression	Glassbox Model
SHAP Kernel Explainer	Black-box Model
LIME	Black-box Model
Morris Sensitivity	Black-box Model
Partial Dependence	Black-box Model



Predicted (2): 1.000 | Actual (2): 1.000



DecisionListClassifier_0 (Overall)

All Rules

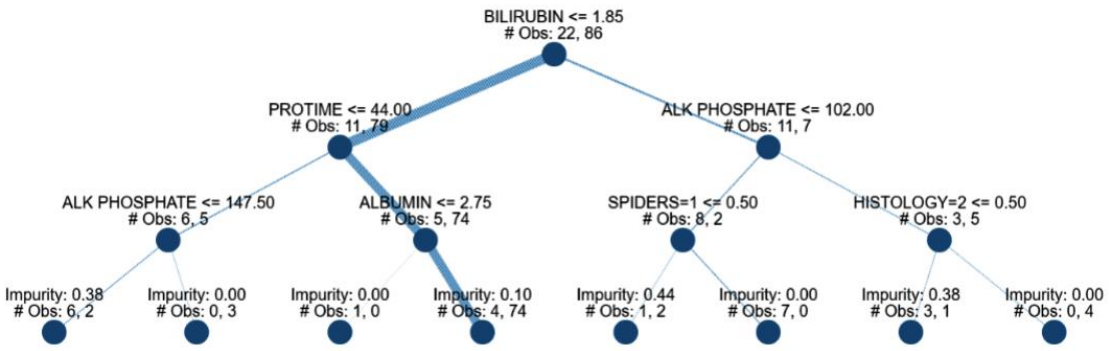
SPIDERS=1 <= 0.50
ALBUMIN > 2.75
PROTIME > 44.00

Prediction: 2, Precision: 1.000, Recall: 0.681

BILIRUBIN <= 1.45
LIVER FIRM=1 > 0.50
SPIDERS=2 <= 0.50

Prediction: 2, Precision: 1.000, Recall: 0.164

ClassificationTree_0 (Overall)



Interpret ML Dashboard

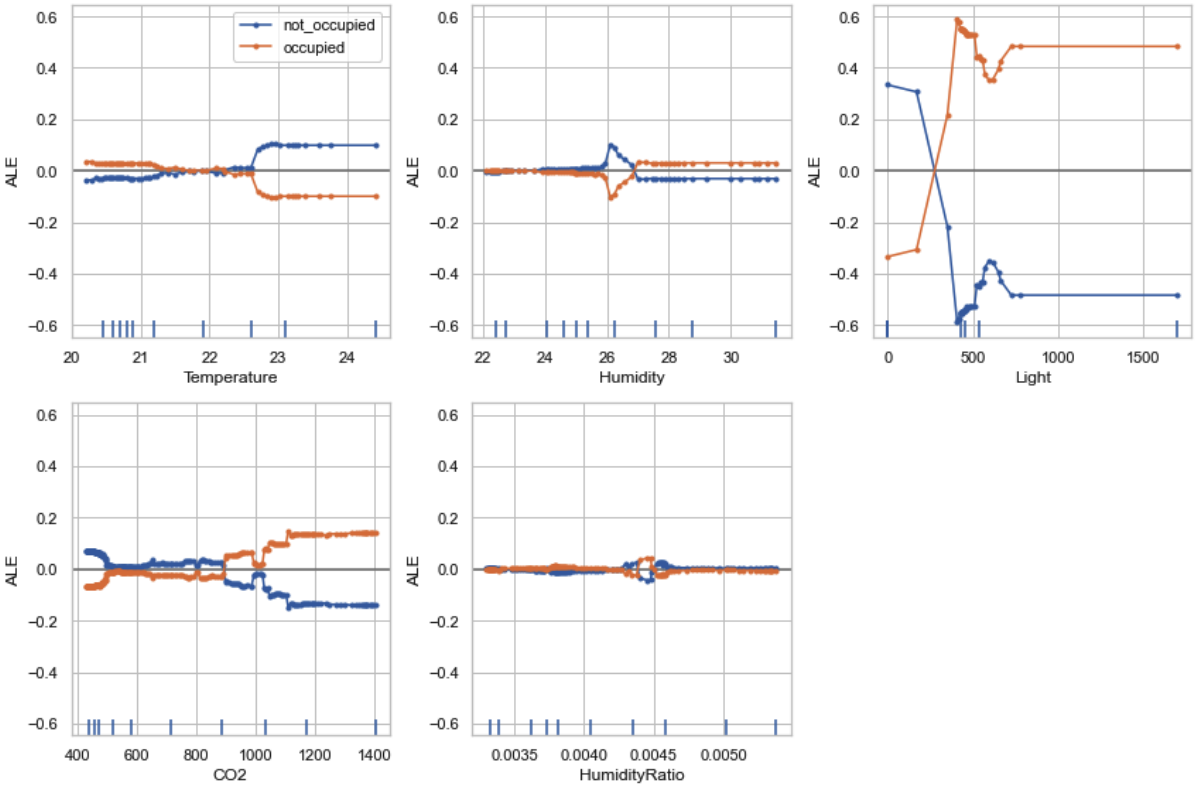
Overview | Data | Performance | Global | Local

Introduction

Welcome to Interpret ML's dashboard. Here you will find en-masse visualizations for your machine learning pipeline.

The explanations available are split into tabs, each covering an aspect of the pipeline.

- **Data** covers exploratory data analysis, designed mostly for feature-level.
- **Performance** covers model performance both overall and user-defined groups.
- **Global** explains model decisions overall.
- **Local** explains a model decision for every instance/observation.

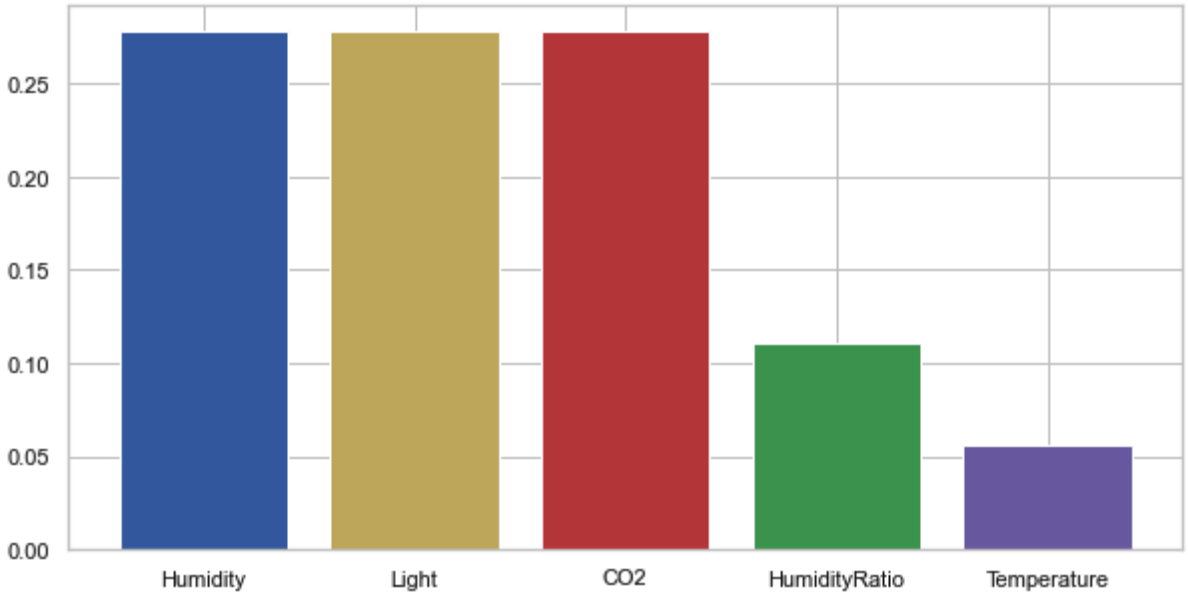


```
Diverse Counterfactuals found! total time taken: 00 min 01 sec
Query instance (original outcome : 0)
```

	Temperature	Humidity	Light	CO2	HumidityRatio	Occupancy
0	20.89	23.1	0.0	499.666667	0.003523	0

```
Diverse Counterfactual set (new outcome: 1.0)
```

	Temperature	Humidity	Light	CO2	HumidityRatio	Occupancy
0	-	24.000000000000014	763.5	696.3	-	1.0
1	-	24.000000000000014	1083.9	500.0666667000001	-	1.0
2	-	24.000000000000014	1046.8	500.0666667000001	0.005247647	1.0
3	-	24.000000000000014	908.6	500.0666667000001	0.004907888	1.0



Weight	Feature
0.1273 ± 0.2324	BILIRUBIN
0.0926 ± 0.2142	ALBUMIN
0.0857 ± 0.1815	ALK PHOSPHATE
0.0801 ± 0.1705	PROTIME
0.0653 ± 0.1381	SGOT
0.0629 ± 0.1471	AGE
0.0487 ± 0.1553	SPIDERS=1
0.0477 ± 0.1555	SPIDERS=2
0.0351 ± 0.1186	LIVER FIRM=1
0.0343 ± 0.1298	LIVER FIRM=2
0.0262 ± 0.1058	HISTOLOGY=2
0.0248 ± 0.1046	SPLEEN PALPABLE=1
0.0245 ± 0.1157	ASCITES=1
0.0233 ± 0.1078	ASCITES=2
0.0225 ± 0.0909	MALAISE=1
0.0210 ± 0.0880	SPLEEN PALPABLE=2
0.0205 ± 0.0844	HISTOLOGY=1
0.0194 ± 0.0882	MALAISE=2
0.0161 ± 0.0676	ANOREXIA=1
0.0144 ± 0.0636	STEROID=1
… *12 more* …	

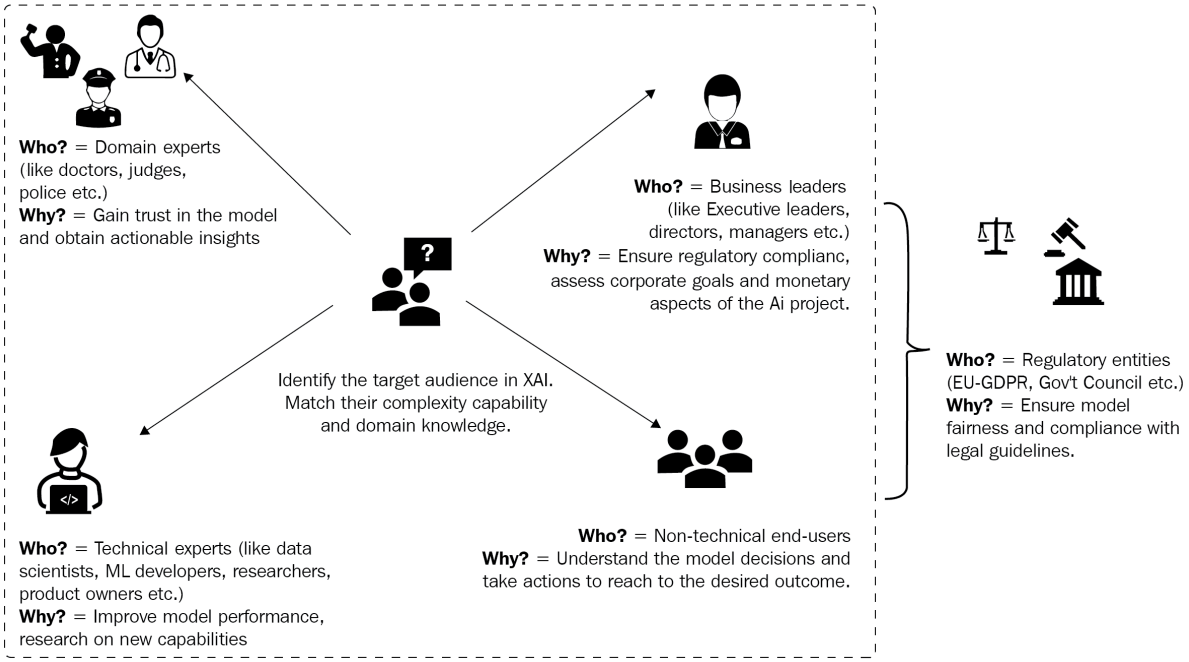
y=Live (probability **0.984**) top features

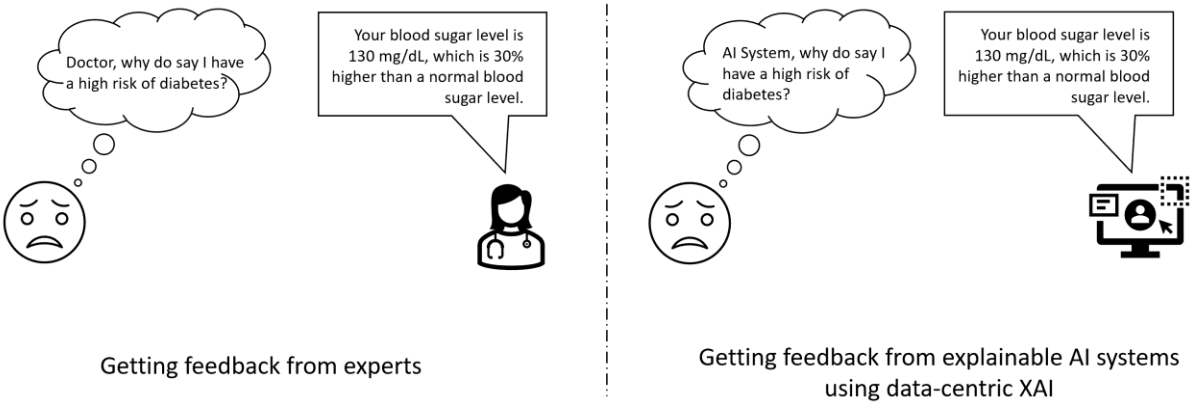
Contribution[?]	Feature	Value
+0.801	<BIAS>	1.000
+0.030	SPIDERS=2	1.000
+0.027	SPIDERS=1	0.000
+0.021	BILIRUBIN	0.700
+0.017	ALBUMIN	4.200
+0.010	HISTOLOGY=2	0.000
+0.009	MALAISE=1	0.000
+0.008	MALAISE=2	1.000
+0.008	FATIGUE=1	0.000
… *19 more positive* …		
… *4 more negative* …		
-0.016	SGOT	224.000

	DALEX	**Explainerdashboard**	**InterpretML**	**ALIBI**	**DiCE**	**ELI5**	**H2O AutoML Explainer**
Data-Centric Explainability	Yes, but not for comprehension	No support	Yes, but very limited option	No support	No support	No support	No support
Model-Centric Explainability	- Supports local and global explanations - Compares model performance and fairness - Supports interactive visualizations	- Supports local and global explanations - Compares model performance - Supports interactive visualizations	- Supports local and global explanations - Compares model performance - Supports interactive visualizations - But limited explainability methods	- Supports local and global explanations but with limited methods	- Supports local and global explanations but only with counterfactuals	- Supports local and global explanations but only with weight-based feature contributions	- Supports local and global explanations - Compares model performance
Outcome-Centric Explainability	Supports prediction-level explanations using What-If analysis and interactive breakdown plots	Supports prediction-level explanations using What-If analysis	No support	No support	No support	No support	Limited support using ICE plots
Compatibility	Highly compatible with scikit-learn, TensorFlow, PyTorch, and H2O models	Only compatible with scikit-learn model formats	Only compatible with scikit-learn model formats	Compatible with scikit-learn and TensorFlow models	Compatible with scikit-learn and TensorFlow, and PyTorch models	Only compatible with scikit-learn models	Only compatible with H2O models
Human-Friendly Explanation	- High for technical users, especially with the interactive dashboard - Low for non-technical users	- High for technical users, due to interactive dashboards - Low for non-technical users	- Not very human friendly in spite of consolidated dashboard	- Not very human friendly	- CFEs are very close to ideal human-friendly explanations but lack interactive visualizations	- Not very human-friendly	- Useful for mostly technical users
Robustness	Considers different dimensions of explanability, so quite robust	Robustness is high as it provides comprehensive model-centric explanations	Not very robust	Not very robust	CFEs can have a lot of randomness and non-feasible suggestions, so robustness is low	Not very robust	Robustness is high as it provides comprehensive model-centric explanations
Scalability	Seems to be scalable and implemented in a production system	Seems to be scalable and implemented in a production system	Not very scalable	Not very scalable	Seems to be scalable and implemented in a production system	Not very scalable	Scalable for H2O models

Chapter 10: XAI Industry Best Practices

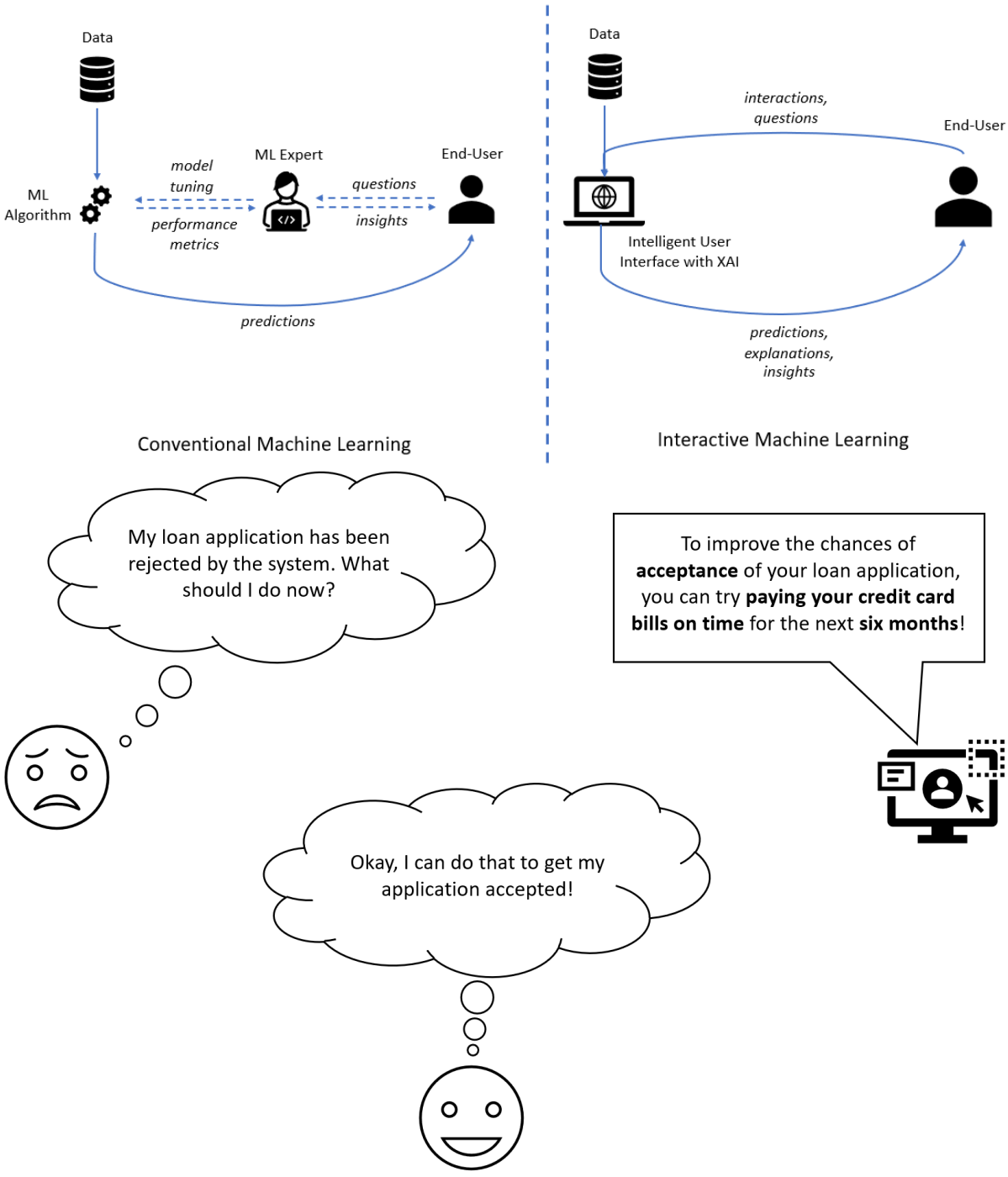






Getting feedback from experts

Getting feedback from explainable AI systems using data-centric XAI


Data
ML Algorithm
model tuning
performance metrics
ML Expert
questions
insights
End-User
predictions
Conventional Machine Learning
Data
interactions, questions
End-User
Intelligent User Interface with XAI
predictions, explanations, insights
Interactive Machine Learning
My loan application has been rejected by the system. What should I do now?
To improve the chances of **acceptance** of your loan application, you can try **paying your credit card bills on time** for the next **six months**!
Okay, I can do that to get my application accepted!

Chapter 11: End User-Centered Artificial Intelligence

